IAN STEWART

WIELKIE

PROBLEMY

MATEMATYCZNE

Prószyński i S-ka

IAN STEWART

WIELKIE PROBLEMY MATEMATYCZNE

Przełożyli
Bogumił Bieniok i Ewa L. Łokas

Prószyński i S-ka

Tytuł oryginału
THE GREAT MATHEMATICAL PROBLEMS
Marvels and Mysteries of Mathematics

Copyright © Joat Enterprises, 2013
First published in Great Britain in 2013 by Profile Books Ltd
All rights reserved

Projekt okładki
© boldandnoble.com

Redaktor prowadzący
Adrian Markowski

Redakcja
Anna Kaniewska

Korekta
Anna Kaniewska

ISBN 978-83-7961-777-7

Warszawa 2014

Wydawca
Prószyński Media Sp. z o.o.
02-697 Warszawa, ul. Rzymowskiego 28
www.proszynski.pl

Musimy wiedzieć. Będziemy wiedzieć.
David Hilbert
w wystąpieniu poświęconym zagadnieniom matematycznym, które wygłosił w 1930 roku na uroczystości nadania mu honorowego obywatelstwa Królewca[1]

1 Te słynne słowa, w oryginale: „Wir müssen wissen. Wir werden wissen", są częścią przemówienia, które Hilbert zarejestrował dla jednej z rozgłośni radiowych. Zob. Constance Reid, *Hilbert*, Springer, Berlin 1970, s. 196.

Przedmowa

Matematyka jest obszerną dziedziną, która bezustannie powiększa się i zmienia. Wśród niezliczonych pytań, jakie stawiają sobie matematycy, są również takie, które szczególnie się wyróżniają – są niczym szczyty górujące nad otaczającymi je pagórkami. Są to pytania tak wielkie, tak trudne i wymagające, że każdy matematyk bez wahania dałby sobie uciąć prawą rękę, gdyby mógł dzięki temu znaleźć na nie odpowiedź. Niektóre pozostawały tajemnicą przez dziesięciolecia, inne przez stulecia, a nieliczne – przez całe tysiąclecia. Niektórych wciąż nie udało się rozwikłać. Wielkie twierdzenie Fermata pozostawało zagadką przez 350 lat i dopiero Andrew Wiles zdołał się z nim uporać po siedmiu latach żmudnej pracy. Hipotezy Poincarégo nikt nie potrafił udowodnić przez ponad sto lat – dokonał tego ekscentryczny geniusz Grigorij Perelman, który odmówił jednak przyjęcia naukowych wyróżnień i nagrody w wysokości miliona dolarów. Hipoteza Riemanna wciąż nie przestaje intrygować matematyków i po 150 latach pozostaje tak samo zagadkowa jak w chwili, gdy ją sformułowano.

Książka *Wielkie problemy matematyczne* zawiera wybór naprawdę wielkich pytań, które sprawiły, że matematyka zaczęła się rozwijać w zupełnie nowych kierunkach. Dowiemy się z niej, w jaki sposób matematycy doszli do tych zagadnień i dlaczego są one tak ważne, poznamy też ich matematyczny i naukowy kontekst. Książka zawiera problemy już rozwiązane i te, z którymi wciąż nie udało się nam uporać. Takie zagadnienia formułowano w różnych okresach w ciągu dwóch tysięcy lat historii rozwoju matematyki, jednak w tej książce skupimy się na problemach, które wciąż pozostają bez odpowiedzi lub zostały rozwiązane w minionym półwieczu.

Podstawowym celem matematyki jest odkrywanie prostoty leżącej u podstaw pozornie skomplikowanych problemów. Jednak nie zawsze

jest to od razu widoczne, ponieważ w matematycznym ujęciu pojęcie „prostoty” bazuje na wielu specjalistycznych i skomplikowanych zagadnieniach. Dużą zaletą tej książki jest to, że podkreśla ową głęboką prostotę, unikając wszelkich złożoności – a przynajmniej wyjaśnia je za pomocą zrozumiałych pojęć.

Matematyka jest bardziej nowatorska i różnorodna, niż się zwykle sądzi. Z grubsza rzecz biorąc, można przyjąć, że obecnie na całym świecie badania prowadzi około stu tysięcy matematyków, którzy każdego roku publikują ponad *dwa miliony* stron artykułów naukowych poświęconych tej dziedzinie. Nie chodzi tu o jakieś „nowe liczby”, bo matematyka wcale nie tym się zajmuje. Nie są to też „nowe obliczenia”, przypominające jakieś wykonane już wcześniej, tylko nieco większe – choć należy przyznać, że w naszej pracy często musimy przeprowadzać całkiem pokaźne rachunki. Niedawno zespół około 25 matematyków przeprowadził badania z dziedziny algebry, które wymagały wykonania „obliczeń dorównujących rozmiarem Manhattanowi”. Nie jest to do końca prawdą, ale błąd polega w tym wypadku raczej na zbyt ostrożnym opisie złożoności problemu. W istocie należałoby powiedzieć, że to *odpowiedź* miała rozmiar Manhattanu – same obliczenia były znacznie większe. To robi wrażenie, ale tak naprawdę liczy się jakość, a nie ilość. Wspomniane obliczenia o rozmiarze Manhattanu są jednak również ważne ze względu na swoją zawartość, ponieważ dostarczają cennych podstawowych informacji na temat grup symetrii, które odgrywają istotną rolę w fizyce kwantowej i matematyce. Genialne odkrycie matematyczne może zmieścić się w jednej linii lub wypełnić całą encyklopedię – wszystko zależy od tego, czego wymaga dane zagadnienie.

Gdy myślimy o matematyce, zwykle wyobrażamy sobie grube księgi wypełnione gęsto symbolami i wzorami. Jednak wspomniane dwa miliony stron zawierają więcej słów niż symboli. Słowa są potrzebne, by wyjaśnić kontekst zagadnienia, omówić przebieg argumentacji, znaczenie obliczeń i wyjaśnić, jak to wszystko wpasowuje się w nieustannie rozrastającą się strukturę matematyki. Wielki Carl

Friedrich Gauss zauważył około roku 1800, że istotą matematyki są „pojęcia, a nie równania”. Idee, a nie symbole. To prawda, ale faktem jest, że matematyczne idee wyraża się najczęściej za pomocą symboli. Wiele artykułów naukowych zawiera więcej symboli niż słów. Wzory pozwalają na uzyskanie takiej dokładności wyrażania myśli, jaką trudno byłoby osiągnąć za pomocą słów.

Nierzadko można jednak wyjaśnić matematyczne idee bez użycia wielu symboli. Książka *Wielkie problemy matematyczne* jest przykładem właśnie takiego podejścia. Objaśnia, czym zajmują się matematycy, w jaki sposób rozumują i dlaczego ich dziedzina jest ciekawa i ważna. Co istotne, pokazuje też, w jaki sposób dzisiejsi matematycy stawiają czoło wyzwaniom rzuconym przez poprzednie pokolenia uczonych i wykorzystując dostępne obecnie potężne techniki obliczeniowe, rozwiązują po kolei wielkie zagadki przeszłości – zmieniając przy okazji samą matematykę i nauki ścisłe. Matematyka jest jednym z największych osiągnięć ludzkości i jej wielkie problemy, rozwiązane i nierozwiązane, już od tysiącleci są siłą napędową leżącą u podstaw jej zdumiewającej mocy – i bez wątpienia będą pobudzały jej rozwój jeszcze przez kolejne tysiąclecia.

Coventry, czerwiec 2012 roku

Autorzy ilustracji

Ryc. 31 – http://random.mostlymaths.net

Ryc. 33 – Carles Simó. Ilustracja pochodzi z książki *European Congress of Mathematics, Budapest 1996* (*Europejski kongres matematyczny, Budapeszt 1996*), „Progress in Mathematics" tom 168, Birkhäuser, Bazylea.

Ryc. 43 – Pablo Mininni.

Ryc. 46 – University College, Cork, Irlandia.

Ryc. 50 – Wolfram MathWorld.

1. Wielkie wyzwania

Na antenach stacji telewizyjnych dość rzadko goszczą programy poświęcone matematyce, a dobre audycje tego typu są już zupełną rzadkością. Jedną z najlepszych, jeśli chodzi o wywołanie zainteresowania widzów, a także zawartość merytoryczną, była audycja poświęcona wielkiemu twierdzeniu Fermata. Przygotował ją John Lynch na zamówienie telewizji BBC i została ona wyemitowana w 1996 roku jako jeden z odcinków doskonałej serii programów popularnonaukowych zatytułowanej *Horizon* (*Horyzont*). Simon Singh, który również brał udział w realizacji programu, napisał później na podstawie tej opowieści wspaniałą książkę[2]. Na jednej ze stron internetowych wyznał, że zdumiewający sukces tego programu był dla niego zaskoczeniem:

> Przez 50 minut matematycy mówili o matematyce, trudno więc było się spodziewać, że będzie to hit telewizyjny, ale powstał program, który wzbudził zainteresowanie widzów i zyskał wysokie oceny krytyków. Audycja zdobyła nagrodę BAFTA w kategorii programów dokumentalnych, Prix Italia oraz inne międzynarodowe nagrody, a także była nominowana do nagród Emmy – wszystko to dowodzi, że matematyka może wzbudzać takie same emocje i być równie pasjonująca jak każda inna dziedzina nauki rozwijana na naszej planecie.

Moim zdaniem jest kilka przyczyn sukcesu tego programu telewizyjnego i opartej na nim książki, i wszystkie one są ważne ze względu na historie, które będziemy tu omawiać. Aby zbytnio nie zagmatwać wywodu, skupmy się na analizie samego programu telewizyjnego.

Wielkie twierdzenie Fermata jest jednym z naprawdę wielkich matematycznych problemów, którego źródłem była pozornie niewinna uwaga pozostawiona przez jednego z czołowych matematyków XVII stulecia na marginesie klasycznego podręcznika. O problemie tym zrobiło się głośno, ponieważ nikt nie potrafił dowieść tego, co Pierre de Fermat stwierdził w swojej notatce, i taki stan rzeczy utrzymywał się przez ponad trzysta lat, mimo że wielu niezwykle utalentowanych uczonych nie szczędziło wysiłków, by uporać się z tym wyzwaniem. Nic więc dziwnego, że gdy w 1995 roku brytyjski matematyk Andrew Wiles przedstawił w końcu dowód tego twierdzenia, nikt nie miał wątpliwości, że jest to niezwykłe dokonanie. Nie trzeba było nawet wiedzieć, na czym polega ów problem, nie mówiąc już o zrozumieniu jego rozwiązania. Był to matematyczny odpowiednik zdobycia Mount Everestu.

Oprócz jego niewątpliwego znaczenia dla matematyki osiągnięcie Wilesa jest również niezwykle interesujące z czysto ludzkiego punktu widzenia. W wieku dziesięciu lat Wiles tak się zainteresował wielkim twierdzeniem Fermata, że postanowił zostać matematykiem i znaleźć jego dowód. Zrealizował pierwszą część swojego planu i został specjalistą od teorii liczb, czyli szerokiej dziedziny matematyki, do której należy twierdzenie Fermata. Jednak w miarę jak coraz lepiej poznawał prawdziwą matematykę, cel, jaki przed sobą postawił, zaczął mu się wydawać coraz bardziej nieosiągalny. Wielkie twierdzenie Fermata było dziwną ciekawostką, niezwiązanym z niczym stwierdzeniem, które praktycznie każdy teoretyk liczb mógłby wymyślić na poczekaniu, pod warunkiem że nie wymagałoby się od niego przedstawienia jakiegokolwiek przekonującego dowodu. W żaden sposób nie pasowało do istniejącego arsenału technik obliczeniowych. W liście do Heinricha Olbersa wielki Gauss odrzucił je z miejsca, stwierdzając, że problem ten wydaje mu się „niezbyt interesujący, ponieważ bez trudu można sformułować wiele podobnych propozycji, których nie sposób ani udowodnić, ani obalić”[3]. Wiles doszedł do wniosku, że jego marzenie z dzieciństwa jest niemożliwe do spełnienia, i postanowił odłożyć na razie twierdzenie Fermata na bok. Wtedy jednak, zupełnie

niespodziewanie, inni matematycy dokonali przełomu, który połączył twierdzenie Fermata z głównym nurtem teorii liczb, i tak się akurat złożyło, że Wiles był już ekspertem w tej konkretnej dziedzinie. Gauss, co do niego zupełnie niepodobne, nie docenił znaczenia tego problemu i nie dostrzegł, że można go połączyć z ważnym, choć pozornie odległym obszarem matematyki.

Po odkryciu tego związku Wiles mógł zająć się próbą rozwiązania zagadki Fermata i *jednocześnie* prowadzić ważne badania z zakresu współczesnej teorii liczb. Była to komfortowa sytuacja: gdyby prace nad twierdzeniem Fermata spełzły na niczym, to i tak wszystko, co odkryłby, próbując je udowodnić, nadawałoby się do publikacji. Wiles powrócił więc do problemu Fermata i zajął się nim z ogromnym zapałem. Po siedmiu latach niestrudzonych badań, które prowadził sam, w tajemnicy przed innymi – w matematyce to dość niezwykłe środki ostrożności – doszedł do przekonania, że w końcu znalazł rozwiązanie. Na prestiżowych konferencjach poświęconych teorii liczb wygłosił serię wykładów pod niejasnym tytułem, który jednak nikogo nie zmylił[4]. W środkach masowego przekazu i w środowisku naukowym gruchnęła sensacyjna wiadomość: wielkie twierdzenie Fermata zostało udowodnione.

Dowód był imponujący i elegancki, pełen wspaniałych pomysłów. Niestety, specjaliści szybko wykryli w nim poważną lukę. W historii zmagań z wielkimi nierozwiązanymi problemami matematycznymi taki rozwój wypadków jest na nasze nieszczęście dosyć częsty i na ogół kończy się fatalnie. Tym razem jednak los okazał się łaskawy. Z pomocą swojego byłego studenta Richarda Taylora Wilesowi udało się zapełnić wykrytą lukę, poprawić dowód i przedstawić pełne rozwiązanie. Wiązały się z tym olbrzymie emocje, co bez trudu można było dostrzec w programie telewizyjnym: był to chyba jedyny przypadek w historii, gdy matematyk rozpłakał się przed kamerą, wspominając traumatyczne wydarzenia i ostateczny tryumf.

Być może zauważyliście, że wciąż jeszcze nie powiedzieliśmy, *jak brzmi* owo wielkie twierdzenie Fermata? To świadomy zabieg – zajmiemy się tym w odpowiednim czasie. Nie ma to większego znaczenia dla analizy źródeł sukcesu wspomnianego programu

telewizyjnego. W istocie nawet sami matematycy nie interesowali się nigdy zbytnio tym, czy twierdzenie, które Fermat zanotował na marginesie czytanej książki, jest prawdziwe, czy nie, ponieważ nie zależy od niego żadne ważne zagadnienie matematyczne. Skąd więc całe to zamieszanie? Ponieważ niezwykle istotne w tym wszystkim było to, że społeczność matematyków nie potrafi *znaleźć* rozwiązania. Nie chodzi tu tylko o ujmę na honorze – taki fakt oznacza, że istniejące teorie mają jakieś braki o niebagatelnym znaczeniu. Poza tym sformułowanie tego twierdzenia jest bardzo łatwe, co jeszcze bardziej potęguje wrażenie tajemniczości. Jak to możliwe, że coś z pozoru tak prostego okazało się tak skomplikowane?

Chociaż udowodnienie tego twierdzenia nie miało dla matematyków większego znaczenia, to bardzo ich niepokoiło, że nie potrafią tego dokonać. Jeszcze bardziej zależało im na znalezieniu metody na udowodnienie twierdzenia Fermata, ponieważ musiałaby ona rzucić nowe światło nie tylko na problem postawiony przez Fermata, ale i na wiele innych zagadnień. Z taką sytuacją mamy bardzo często do czynienia w wypadku wielkich problemów matematycznych: to *metody* ich rozwiązania, a nie same rozwiązania, okazują się najważniejsze. Oczywiście czasami same rozwiązania również są ważne – wszystko zależy od tego, jakie wynikają z nich konsekwencje.

Dowód Wilesa najeżony jest trudnymi pojęciami i zbyt skomplikowany, by nadawał się do telewizji. Prawdę mówiąc, szczegóły jego wywodu mogą zrozumieć tylko specjaliści[5]. Z dowodem tym wiąże się interesująca matematyczna opowieść, o czym przekonamy się w swoim czasie, ale wszelkie próby wyjaśnienia tego w telewizji byłyby z góry skazane na porażkę. Zamiast tego twórcy programu słusznie skupili uwagę na bardziej ludzkim aspekcie tych wydarzeń: jak to jest, gdy człowiek zmaga się z trudnym problemem matematycznym, za którym ciągnie się olbrzymi bagaż historii? Widzowie dowiedzieli się, że istnieje niewielka, ale niezwykle oddana sprawie grupka matematyków rozsianych po całym świecie, którym bardzo zależy na rozwoju ich

obszaru badań. Uczeni ci kontaktują się ze sobą, z uwagą czytają prace kolegów i znaczną część swojego życia poświęcają rozwojowi wiedzy matematycznej. W programie świetnie pokazano, ile uczucia wkładają w swoją pracę i jak silne wytwarzają się między nimi więzi społeczne. Wcale nie są inteligentnymi automatami, ale prawdziwymi ludźmi, którzy z oddaniem rozwijają swoją dziedzinę. Taki przekaz popłynął z ekranu.

Oto te trzy ważne przyczyny, dzięki którym program odniósł tak duży sukces: istotny problem, główny bohater ze wspaniałą, ludzką historią i drugoplanowi bohaterowie oddani swojej pracy. Podejrzewam jednak, że znaczenie miał tutaj jeszcze czwarty, mniej krzepiący czynnik. Większość ludzi niezwiązanych z matematyką bardzo rzadko dowiaduje się o najnowszych osiągnięciach z tej dziedziny. Przyczyny tego są różne i wszystkie całkowicie zrozumiałe: ludzie i tak niespecjalnie się tym interesują; w gazetach rzadko kiedy wspomina się o czymkolwiek, co ma związek z matematyką, a jeśli już do tego dochodzi, to doniesienia te są często niepoważne lub trywialne. Poza tym wydaje się, że nic w życiu zwyczajnego człowieka nie ma związku z tym, co robią matematycy w zaciszu swoich gabinetów. W szkole zbyt często matematykę przedstawia się jako zamkniętą księgę, w której każde pytanie ma swoją odpowiedź. Uczniowie dochodzą więc zazwyczaj do wniosku, że nowe odkrycia matematyczne są równie rzadkie jak białe kruki.

Z takiego punktu widzenia interesującą wiadomością nie było to, że udowodniono wielkie twierdzenie Fermata. Sensacja polegała na tym, że w końcu *ktoś dokonał w matematyce jakiegoś nowego odkrycia*. Ponieważ matematycy potrzebowali ponad trzystu lat, żeby znaleźć ten dowód, widzowie podświadomie przyjęli, że przełom ten był pierwszym ważnym odkryciem matematycznym od trzystu lat. Nie twierdzę, że świadomie w to wierzyli. Takie stwierdzenie jest trudne do obrony, gdy tylko zastanowimy się nad kilkoma oczywistymi kwestiami, takimi jak: „Dlaczego rząd wydaje pieniądze na uniwersyteckie wydziały matematyki?”. Jednak wiele osób przyjęło podświadomie takie właśnie założenie, nie zastanawiając się nad tym ani nie analizując jego zasadności. Dzięki temu osiągnięcie Wilesa wydało się jeszcze większe.

Jednym z celów tej książki jest pokazanie, że matematyka jest prężnie rozwijającą się dziedziną nauki, w której nieustannie dokonuje się nowych odkryć. Nie słyszy się zbyt wiele o tych dokonaniach, ponieważ większość z nich jest zbyt skomplikowana, by mogły je zrozumieć osoby niezajmujące się matematyką, a media jak diabeł święconej wody boją się wszystkiego, co wymagałoby od widzów większego wysiłku intelektualnego niż *Taniec z gwiazdami*. Poza tym rozmyślnie ukrywa się wszelkie zastosowania matematyki, żeby nikogo niepotrzebnie nie niepokoić: „Co takiego? Działanie mojego iPhone'a opiera się na wykorzystaniu zaawansowanej matematyki? No to jak się zaloguję na Facebooka, skoro nie zdałem egzaminu z matematyki?".

Jak uczy nas historia, matematyka często rozwija się w wyniku odkryć dokonywanych w innych dziedzinach. Gdy Newton pracował nad zasadami dynamiki i prawem powszechnego ciążenia, które razem pozwoliły nam opisać ruch planet, wcale nie starał się do końca zrozumieć Układu Słonecznego. Przeciwnie, to matematycy musieli stawić czoło całej nowej kategorii pytań i zastanowić się, co w istocie wynika z tych nowych praw. Aby odpowiedzieć na to pytanie, Newton wymyślił rachunek różniczkowy i całkowy, ale jego nowa metoda ma swoje ograniczenia. Często prowadzi jedynie do uzyskania innego sformułowania postawionego pytania, a nie odpowiedzi. Przekształca problem w szczególny rodzaj wzoru, zwanego równaniem różniczkowym, którego *rozwiązanie* jest szukaną odpowiedzią. Zatem wciąż jeszcze trzeba rozwiązać to równanie. Niemniej rachunek różniczkowy i całkowy był doskonałym punktem wyjścia. Newton pokazał, że uzyskanie odpowiedzi jest możliwe, i przedstawił skuteczną metodę ich poszukiwania, która trzysta lat później wciąż dostarcza nam cennych informacji.

W miarę jak powiększała się sumaryczna wiedza matematyczna ludzkości, coraz większą rolę w rozwoju tej dziedziny zaczęło odgrywać drugie źródło inspiracji: wewnętrzne potrzeby samej matematyki. Jeśli na przykład wiemy, jak rozwiązuje się równania algebraiczne pierwszego, drugiego, trzeciego i czwartego stopnia, to

nie trzeba zbyt dużej wyobraźni, żeby zapytać o równania stopnia piątego. (Stopień równania jest w zasadzie miarą jego złożoności, ale nie trzeba tego wcale wiedzieć, żeby zadać to oczywiste pytanie). Jeśli uzyskanie odpowiedzi na takie pytanie okazuje się trudne – a tak było w tym wypadku – to fakt ten *sam w sobie* pobudza matematyków do szukania rozwiązania z jeszcze większym zaangażowaniem, bez względu na to, czy znajdzie ono jakieś praktyczne zastosowanie.

Nie chcę przez to powiedzieć, że zastosowania praktyczne nie mają znaczenia. Jeżeli jednak jakieś określone zagadnienie matematyczne pojawia się bezustannie w rozważaniach związanych z fizyką fal – fal morskich, drgań, dźwięku, światła – to bez wątpienia warto taki obszar zbadać choćby tylko po to, by lepiej go poznać. Nie musimy z góry wiedzieć, w jaki sposób dana nowa idea zostanie wykorzystana: fale występują w tak wielu ważnych dziedzinach nauki, że z pewnością każde istotne odkrycie z nimi związane do czegoś się w końcu przyda. W tym konkretnym przypadku odkrycia przydały się w technice radiowej, telewizyjnej i radarowej[6]. Jeśli natomiast ktoś wymyśli jakiś nowy sposób rozumienia przepływu ciepła i wpadnie na genialny pomysł, któremu będzie niestety brakowało odpowiedniego wsparcia matematycznego, to najsensowniej będzie uporać się z tym problemem, analizując go *jako część matematyki*. Ktoś, kogo nie interesuje ani trochę, jak przepływa ciepło, zawsze może przyjąć, że uzyskane wyniki znajdą zapewne jakieś inne zastosowanie. Transformacja Fouriera, która narodziła się w wyniku takich właśnie badań, jest obecnie najbardziej chyba użyteczną ideą matematyczną, jaką kiedykolwiek wymyślono. Leży u podstaw współczesnej telekomunikacji, dzięki niej działają aparaty cyfrowe, umożliwia oczyszczenie z szumów starych filmów i nagrań, a FBI wykorzystuje jej współczesne rozszerzenie do przechowywania odcisków palców[7].

Taka wzajemna wymiana idei między zewnętrznymi zastosowaniami matematyki a jej wewnętrzną strukturą trwa od kilku tysiącleci i oba te aspekty splotły się ze sobą tak ściśle, że ich rozdzielenie jest już w zasadzie niemożliwe. Bez większych trudności

możemy jednak rozróżnić nasze nastawienie do matematyki, co prowadzi do szerokiego jej podziału na dwa rodzaje: teoretyczna i stosowana. Taki podział ma rację bytu jako prosty sposób na umiejscowienie idei matematycznych w krajobrazie pojęciowym, ale nie jest zbyt dokładnym opisem samej matematyki. W najlepszym wypadku pozwala na odróżnienie dwóch krańców szerokiego, ciągłego zakresu matematycznych podejść. W najgorszym – prowadzi do nieporozumień odnośnie do tego, które gałęzie matematyki są użyteczne i skąd pochodzą dane idee. Podobnie jak to się dzieje we wszystkich dziedzinach nauki, również matematyka czerpie swoją moc z *połączenia* abstrakcyjnego rozumowania z ideami płynącymi ze świata zewnętrznego – i oba te aspekty wzajemnie się inspirują. Rozdzielenie obydwu wątków nie tylko jest niemożliwe – działanie takie jest bezcelowe.

Większość naprawdę ważnych problemów matematycznych, owe wielkie wyzwania, którym poświęcona jest ta książka, powstała w ramach samej matematyki, w wyniku swego rodzaju intelektualnego zapatrzenia we własny pępek. Przyczyna tego jest prosta: są to problemy *matematyczne*. Matematykę często postrzega się jako zbiorowisko odrębnych działów, z których każdy ma własne specjalistyczne techniki: algebra, geometria, trygonometria, analiza, kombinatoryka, probabilistyka. Tak się jej też naucza i są po temu dobre powody: umiejscowienie każdego oddzielnego zagadnienia w jednym, dobrze określonym obszarze pozwala uczniom uporządkować materiał w głowie. Jest to sensowne pierwsze przybliżenie struktury matematyki, szczególnie uzasadnione dla działów od dawna już zbadanych. Jednak na obszarach toczących się obecnie badań naukowych taki wyraźny podział ulega często rozmyciu. I nie chodzi tu tylko o to, że granice między głównymi obszarami matematyki zamazują się – one po prostu nie istnieją.

Każdy matematyk prowadzący badania ma pełną świadomość tego, że w każdej chwili, nagle i niespodziewanie, problem, nad którym pracuje, może wymagać zastosowania idei z jakiegoś innego, pozornie niezwiązanego z nim działu. Nowe badania w istocie prowadzą często do połączenia różnych gałęzi matematyki. Na przykład moje skupiają się wokół zagadnienia powstawania wzorów

w układach dynamicznych, czyli takich, które zmieniają się z upływem czasu zgodnie z określonymi zasadami. Typowym przykładem jest sposób poruszania się zwierząt. Kłusujący koń powtarza bez końca taki sam ciąg ruchów, co prowadzi do powstania wyraźnego wzorca: kopyta opadają na ziemię przeciwległymi parami, to znaczy – najpierw od ziemi odbijają się nogi lewa przednia i prawa tylna, a potem pozostałe dwie. Czy jest to zagadnienie związane z wzorami? W takim razie oznaczałoby to, że metody jego analizy powinny się wywodzić z teorii grup, czyli algebry symetrii. A może jest to zagadnienie związane z dynamiką? Wówczas należałoby zastosować równania różniczkowe w stylu newtonowskim.

Prawda jest taka, że z definicji tego typu problem jest powiązany z obydwiema dziedzinami. Nie chodzi tu o ich część wspólną, czyli zakres materiału wspólny dla obu dziedzin, bo to jest w zasadzie zbiór pusty. Mówimy tu raczej o nowym „obszarze" łączącym obie tradycyjnie rozdzielne dziedziny. Przypomina on most przerzucony przez rzekę graniczną między dwoma państwami – most taki łączy obydwa kraje, choć nie należy do żadnego z nich. W naszym przykładzie jednak most ten nie jest wąskim pasem drogi – jego rozmiar jest porównywalny z rozmiarem każdego z sąsiadujących państw. Jeszcze ważniejsze jest to, że wykorzystywane w tym wypadku metody nie ograniczają się do technik stosowanych na obydwu obszarach. W istocie każda dziedzina matematyki, której się kiedykolwiek uczyłem, odgrywa w moich badaniach jakąś rolę. Na studiach na uniwersytecie w Cambridge poznałem teorię Galois, która mówi o tym, jak należy rozwiązywać równania algebraiczne piątego stopnia (a właściwie o tym, dlaczego nie można ich rozwiązać). Na zajęciach poświęconych teorii grafów poznałem sieci, czyli punkty połączone liniami. Nigdy nie uczyłem się układów dynamicznych, ponieważ moja praca doktorska była poświęcona algebrze, ale w ciągu tych wszystkich lat poznałem podstawy tej dziedziny, od stanów stacjonarnych do chaosu. Teoria Galois, teoria grafów, układy dynamiczne – trzy oddzielne obszary. Tak przynajmniej sądziłem do 2011 roku, gdy zapragnąłem zrozumieć, jak można wykryć zachowanie chaotyczne w sieci układów dynamicznych, i okazało się, że kluczowy krok wymagał

zastosowania teorii Galois, którą poznałem 45 lat wcześniej na studiach.

Matematyka nie przypomina zatem politycznej mapy świata, na której każdy kraj ma jednoznacznie wytyczoną, wyraźną granicę i jest zaznaczony innym kolorem – różowym, zielonym lub jasnoniebieskim – dzięki czemu szybko można go odróżnić od państw sąsiednich. Bardziej przypomina pejzaż, na którym trudno tak naprawdę stwierdzić, gdzie kończy się dolina, a zaczyna wzgórze, gdzie las przechodzi w zagajnik, a ten z kolei w chaszcze i trawiastą łąkę. W takim pejzażu jeziora rozlewają swe wody na każdym rodzaju terenu, a rzeki łączą okryte śniegiem wierzchołki gór z dalekimi, rozległymi morzami. Jednak takiego wiecznie zmieniającego się matematycznego krajobrazu nie tworzą skały, woda i rośliny, ale idee, a łączącym go spoiwem nie jest geografia, lecz logika. Jest to krajobraz dynamiczny, który zmienia się za każdym razem, gdy ktoś odkrywa jakieś nowe idee lub metody. Ważne pojęcia o szerokich zastosowaniach górują niczym wysokie szczyty, a często używane metody przypominają szerokie rzeki niosące podróżnych przez żyzne równiny. Im dokładniej przyjrzymy się temu pejzażowi, tym łatwiej będziemy mogli dostrzec w nim niezdobyte szczyty lub nieprzebyte obszary tworzące niepotrzebne przeszkody. Z czasem niektóre z tych szczytów i przeszkód zyskują sławę. To są właśnie nasze wielkie wyzwania.

Co sprawia, że problem matematyczny staje się wielki? Problem taki musi się cechować głębią intelektualną w połączeniu z prostotą i elegancją. I jeszcze jedno: musi być *trudny*. Każdy może wejść na pagórek – zdobycie Mount Everestu to zupełnie co innego. Wielki problem można zwykle wyrazić w prostej formie, choć użyte wyrażenia mogą być różne: elementarne lub niezwykle zaawansowane. Zapis wielkiego twierdzenia Fermata lub twierdzenia o czterech barwach jest z miejsca zrozumiały dla każdego, kto uczył się w szkole matematyki. Natomiast hipotezy Hodge'a lub hipotezy luki masowej nie można nawet sformułować bez odwoływania się do skomplikowanych pojęć z najnowszych obszarów badań – nie

przypadkiem hipoteza luki masowej pojawiła się na gruncie kwantowej teorii pola. Jednak dla osób obeznanych z tymi dziedzinami sformułowanie wymienionych zagadnień jest proste i naturalne. Nie wymaga zapisania wielu stron drobnym maczkiem. Gdzieś pośrodku plasują się problemy wymagające znajomości matematyki z zakresu szkoły średniej i wyższej – jeśli ktoś pragnie je zrozumieć dokładnie. Każdy natomiast może pojąć na bardziej ogólnym poziomie najważniejsze aspekty danego problemu – skąd się wziął, dlaczego jest ważny, co uzyskamy dzięki jego rozwiązaniu – i właśnie takie wyjaśnienia będę się starał przedstawić w tej książce. Przyznaję, że hipoteza Hodge'a okazuje się pod tym względem twardym orzechem do zgryzienia, ponieważ jest bardzo skomplikowana i abstrakcyjna. Znalazła się jednak na liście siedmiu matematycznych problemów milenijnych ogłoszonej przez Instytut Claya i za jej udowodnienie wyznaczono nagrodę w wysokości miliona dolarów, nie możemy więc jej tu pominąć.

Wielkie problemy pobudzają nas do twórczego działania – pomagają odkrywać nowe obszary matematyki. W 1900 roku na Międzynarodowym Kongresie Matematyków w Paryżu David Hilbert wygłosił wykład, na którym przedstawił listę 23 najważniejszych problemów matematycznych. Nie umieścił na niej wielkiego twierdzenia Fermata, ale wspomniał o nim na początku wykładu. Gdy jakiś wybitny matematyk wymienia najważniejsze jego zdaniem problemy, pozostali uczeni słuchają go z uwagą. Wymienione zagadnienia nie znalazłyby się na tej liście, gdyby nie były ważne i trudne. To zupełnie naturalne, że takie problemy stają się wyzwaniem rzuconym społeczności uczonych i wszyscy próbują je rozwiązać. Od czasu kongresu w Paryżu rozwiązanie któregoś z problemów Hilberta było doskonałym sposobem na zdobycie matematycznych ostróg. Część z tych problemów jest zbyt skomplikowana, by można je było omówić w tej książce, część ma charakter otwartego programu badawczego, a nie konkretnego zagadnienia do rozwiązania, jednak kilka z nich pojawi się w dalszej części naszej opowieści. Wszystkie zaś zasługują na wymienienie, dlatego w przypisach zamieściłem ich krótkie podsumowanie[8].

Właśnie to sprawia, że dany problem matematyczny jest wielki. Natomiast fakt, że dane zagadnienie staje się problemem, rzadko kiedy oznacza, iż nie wiadomo, jak powinno wyglądać rozwiązanie. Praktycznie w wypadku wszystkich wielkich problemów matematycy mają (lub mieli, jeśli chodzi o problemy już rozwiązane) bardzo dobre wyobrażenie na temat tego, jaka powinna być odpowiedź. Co więcej, samo sformułowanie problemu często zawiera już oczekiwane rozwiązanie. Każda hipoteza tak właśnie wygląda – jest przypuszczalnym twierdzeniem wynikającym z różnorodnych dowodów. Większość dogłębnie zbadanych hipotez okazuje się ostatecznie prawdziwa, ale nie wszystkie. Zdarza się, że uczeni używają innych określeń, mając na myśli hipotezę – na przykład dla twierdzenia Fermata słowo „twierdzenie" jest (a mówiąc dokładnie, było) nadużyciem – twierdzenie wymaga dowodu, a akurat jego brakowało, dopóki nie pojawił się Wiles.

To właśnie konieczność przedstawienia dowodu powoduje, że wielkie problemy zasługują na miano problemów. Każdy średnio wykształcony matematyk może przeprowadzić parę obliczeń, dostrzec pojawiającą się prawidłowość i zawrzeć jej sedno w jakimś mniej lub bardziej nieporadnym stwierdzeniu. Matematycy wymagają silniejszych dowodów – żądają pełnego, nienagannego pod względem logiki dowodu. Albo też, jeśli odpowiedź jest przecząca – dowodu obalającego dane stwierdzenie. Trudno w istocie w pełni docenić nieodparty urok wielkich problemów bez zrozumienia kluczowej roli, jaką odgrywa dowód we wszelkich działaniach matematyków. Każdy może przeanalizować garść faktów i wysunąć hipotezę. Cała trudność polega na udowodnieniu, że jest ona poprawna. Albo błędna.

Znaczenie pojęcia dowodu matematycznego ulegało w przeszłości zmianom, które najczęściej polegały na coraz silniejszym zaostrzaniu wymagań logicznych. Uczeni odbyli wiele filozoficznych dyskusji na temat natury dowodu i dzięki nim udało się zwrócić uwagę na kilka istotnych kwestii. W efekcie zaproponowano i zaczęto stosować w praktyce dokładną logiczną definicję „dowodu". W szkole uczą nas, że dowód zaczyna się od przedstawienia jawnych założeń, które nazywamy aksjomatami. Można powiedzieć, że definiują one reguły

gry. Równie dobrze moglibyśmy przyjąć inne aksjomaty, ale wtedy mówilibyśmy o innej grze. Podejście takie wprowadził starożytny grecki matematyk Euklides i mimo upływu lat wciąż pozostaje ono ważne. Gdy już ustalimy, jakie są aksjomaty, możemy przeprowadzić dowód danego twierdzenia, który polega na przedstawieniu kolejnych kroków będących logicznym wnioskiem płynącym albo z aksjomatów, albo z poprzednich kroków, albo z jednego i drugiego. Można powiedzieć, że matematyk penetruje logiczny labirynt, w którym rolę rozwidleń odgrywają kolejne stwierdzenia, a przejścia to poprawnie wysnute wnioski. Dowód jest w takim wypadku ścieżką prowadzącą przez labirynt, która zaczyna się od aksjomatów. Proces taki dowodzi prawdziwości stwierdzenia, przy którym kończy się ścieżka.

Takie proste pojęcie dowodu nie oddaje jednak całej złożoności zagadnienia. Nie jest nawet najważniejszą częścią całego procesu dowodzenia. To tak, jakby powiedzieć, że symfonia jest ciągiem dźwięków zestawionych zgodnie z regułami harmonii. Takie stwierdzenie całkowicie pomija czynnik twórczy. Nie mówi nic na temat tego, jak należy szukać dowodów ani nawet jak sprawdzić poprawność dowodu, który przeprowadził ktoś inny. Nie wspomina ani słowem, które miejsca w labiryncie są ważne. Nie dowiemy się też, jakie ścieżki są eleganckie, a jakie brzydkie; które są ważne, a które – bez znaczenia. Jest to formalny, mechaniczny opis procesu, który ma wiele różnych aspektów, a w szczególności – wymiar ludzki. To ludzie odkrywają dowody i badania matematyczne nie sprowadzają się jedynie do wykonywania kolejnych kroków wynikających z logiki.

Gdyby potraktować tę formalną definicję dosłownie, uzyskalibyśmy praktycznie nieczytelne dowody, ponieważ większość czasu musielibyśmy poświęcić na stawianie każdej kropki nad każdym logicznym „i”, podczas gdy wynik końcowy od dawna byłby już wiadomy. Dlatego matematycy idą drogą na skróty i pomijają wszystko, co jest oczywiste lub stanowi część rutynowych działań. Zawsze jednak zaznaczają, że w danym miejscu następuje przeskok w wywodzie, używając standardowych stwierdzeń, takich jak „łatwo można stwierdzić, że…” lub „z prostych obliczeń wynika, że”. Jedyne,

czego nie robią, przynajmniej nie świadomie, to nie omijają chyłkiem logicznych trudności ani nie próbują udawać, że ich w danym miejscu nie ma. Ba, doświadczeni matematycy wkładają nawet wiele wysiłku w dokładne pokazanie właśnie tych fragmentów argumentacji, które nie są do końca pewne z logicznego punktu widzenia, i poświęcają większość czasu na wyjaśnienie, co należałoby zrobić, żeby je odpowiednio wzmocnić. W efekcie dowód jest matematyczną opowieścią z własnym przebiegiem wydarzeń. Ma początek, środek i koniec. Często pojawiają się wątki poboczne, wyrastające z głównej myśli, i każdy z nich znajduje swoje rozwiązanie. Brytyjski matematyk Christopher Zeeman zauważył kiedyś, że twierdzenie jest intelektualnym miejscem wytchnienia. Możemy się na chwilę zatrzymać, złapać oddech i poczuć, że dokądś dotarliśmy. Wątek poboczny służy wyjaśnieniu jakiegoś szczegółu głównego toku narracji. Dowody przypominają opowieści również pod innymi względami: często występuje w nich jeden lub kilkoro głównych bohaterów – są to oczywiście idee, nie osoby – i łączące ich skomplikowane związki prowadzą do ostatecznego rozwiązania.

Jak wynika ze szkolnej definicji, dowód rozpoczyna się od jawnego wymienienia założeń, a następnie następuje wyprowadzenie logicznych konsekwencji przedstawione w spójny i uporządkowany sposób, prowadzące do tego, co chcemy udowodnić. Jednak dowód nie jest tylko listą wyciąganych wniosków, a logika nie jest jedynym stosowanym tu kryterium. Dowód jest opowieścią przygotowaną dla ludzi, którzy większość czasu poświęcają na naukę czytania tak zapisanych historii i znajdowania w nich pomyłek lub niespójności. Ich głównym celem jest udowodnienie, że autor *się myli*. Osoby te są obdarzone osobliwym darem dostrzegania wszelkich słabości i z całą bezwzględnością będą uderzały w każdy słaby punkt tak długo, aż cała konstrukcja runie, wzbudzając tumany kurzu. Gdy matematyk dochodzi do wniosku, że udało mu się rozwiązać jakieś istotne zagadnienie – nieważne, czy będzie to wielki problem, czy też coś ciekawego, ale wzbudzającego mniejszy podziw – to w pierwszym odruchu nie krzyczy „Hura!” i nie sięga po butelkę szampana, ale próbuje obalić to, co przed chwilą osiągnął.

Być może brzmi to zniechęcająco, ale dowód jest jedynym wiarygodnym narzędziem matematyków, pozwalającym upewnić się, że to, co mówią, jest prawdziwe. Przewidując taką reakcję kolegów, badacze wkładają wiele wysiłku w próby obalenia własnych pomysłów i dowodów. W ten sposób mogą sobie oszczędzić wstydu. Gdy dana opowieść wychodzi obronną ręką z tego rodzaju krytycznej oceny, uczeni bezzwłocznie uznają, że jest ona poprawna, i wtedy jej autor otrzymuje należne mu słowa uznania, szacunek i nagrodę. A przynajmniej tak to się zazwyczaj odbywa, choć osoby zaangażowane w ten proces mogą czasami odnieść inne wrażenie. Gdy ktoś jest w samym środku akcji, jego ocena przebiegu wydarzeń może odbiegać od tego, co zauważa obserwator patrzący z zewnątrz.

W jaki sposób matematycy rozwiązują problemy? Przeprowadzono kilka rygorystycznych badań naukowych na ten temat. Współczesne badania procesu nauczania, oparte na naukach kognitywnych, skupiają się na procesie edukacji do poziomu szkoły średniej. Niektóre badania obejmują również uczelnie wyższe, ale należą one do rzadkości. Istnieją istotne różnice między uczeniem się i nauczaniem poznanej już matematyki a odkrywaniem nowych obszarów badań matematycznych. Wielu z nas umie grać na instrumencie muzycznym, ale tylko nieliczni potrafią skomponować koncert, czy choćby napisać piosenkę.

Gdy w grę wchodzi proces twórczy na najwyższym poziomie, znakomita większość tego, co wiemy – lub co sądzimy, że wiemy – pochodzi z analizy własnych odczuć i przemyśleń. Prosimy matematyków, żeby wyjaśnili nam swój proces myślowy, i spodziewamy się, że usłyszymy jakieś ogólne zasady. Jedną z pierwszych poważnych prób odkrycia, jak myślą matematycy, była praca Jacques'a Hadamarda z 1945 roku, która ukazała się pod tytułem *Psychologia odkryć matematycznych*. Hadamard przeprowadził rozmowy z czołowymi matematykami i uczonymi swoich czasów, w których prosił ich o to, by opisali, jak myślą, gdy rozwiązują jakiś trudny problem. Z rozmów tych jasno wynika, że kluczową rolę w tym procesie odgrywa coś, co z braku lepszego

słowa możemy nazwać intuicją. Myślami rozmówców Hadamarda kierowała jakaś część podświadomości. Na najlepsze pomysły nie wpadali po przeprowadzeniu rygorystycznego wywodu logicznego, ale w wyniku nagłych, niezwykłych olśnień.

Autorem jednego z najbardziej szczegółowych opisów takiego pozornie nielogicznego podejścia do rozważań logicznych jest francuski matematyk Henri Poincaré, jeden z najważniejszych uczonych końca XIX i początków XX wieku. Poincaré zajmował się zagadnieniami praktycznie ze wszystkich dziedzin matematyki, zapoczątkował kilka nowych obszarów badań i radykalnie odmienił wiele innych. Spotkamy się z nim jeszcze w kilku dalszych rozdziałach tej książki. Był również autorem książek popularnonaukowych i to doświadczenie pomogło mu być może lepiej zrozumieć własny proces myślowy. Tak czy inaczej, Poincaré stanowczo twierdził, że świadome rozważania logiczne są tylko częścią procesu twórczego. Tak, w pewnych momentach są one nieodzowne – gdy trzeba ustalić, na czym właściwie polega problem, czy systematycznie sprawdzić uzyskane rozwiązanie. Jednak w pozostałych chwilach Poincaré miał poczucie, że jego mózg pracuje nad danym zagadnieniem, zupełnie go o tym nie informując, w sposób, którego nie potrafił do końca zrozumieć.

W swoim opisie procesu twórczego wyróżnił trzy kluczowe etapy: przygotowanie, dojrzewanie i olśnienie. Przygotowanie składa się ze świadomych wysiłków logicznych mających na celu uchwycenie problemu, uściślenie go i zmierzenie się z nim z wykorzystaniem standardowych metod. W opinii Poincarégo ten etap ma podstawowe znaczenie: prowadzi do uruchomienia podświadomości i dostarcza jej surowca do pracy. Dojrzewanie rozpoczyna się wtedy, gdy przestajemy rozmyślać o danym problemie i zajmujemy się czymś innym. Podświadomość łączy teraz ze sobą różne idee, często całkiem szalone, tak długo, aż zacznie nam coś świtać. Jeśli dopisze nam szczęście, prowadzi to do olśnienia: podświadomość klepie nas po ramieniu, a w mózgu zapala się przysłowiowa lampka.

Taki proces twórczy przypomina spacer po linie. Z jednej strony nie uda nam się rozwiązać trudnego problemu, jeśli nie poznamy obszaru, do którego zdaje się należeć – a także, tak na wszelki

wypadek, wielu innych obszarów, które nie muszą mieć z nim związku. Z drugiej jednak strony, jeśli ograniczymy się do standardowego sposobu myślenia, podążając śladem wielu innych uczonych, którzy bezskutecznie przemierzali już te okolice, to wpadniemy w myślowe koleiny i nie odkryjemy nic nowego. Sztuczka polega więc na tym, żeby wiedzieć jak najwięcej, powiązać świadomie tę wiedzę ze sobą, łamać sobie głowę nad danym problemem przez całe tygodnie, a potem… odłożyć go na bok. Wtedy do roboty zabiera się intuicyjna część mózgu, która zderza ze sobą różne idee, żeby sprawdzić, czy posypią się iskry, i gdy coś odkryje, bezzwłocznie nas o tym zawiadamia. Może się to zdarzyć w dowolnym momencie: Poincaré męczył się kiedyś przez całe miesiące nad pewnym problemem, a potem niespodziewanie zrozumiał nagle, w chwili gdy wysiadał z autobusu, jak można go rozwiązać. Srinivasa Ramanujan, indyjski matematyk, genialny samouk obdarzony niezwykłym talentem do odkrywania wspaniałych wzorów, często wpadał na doskonałe pomysły podczas snu. Archimedes natomiast, jak wszystkim wiadomo, wymyślił sposób na sprawdzenie, czy dany metal jest złotem, podczas kąpieli.

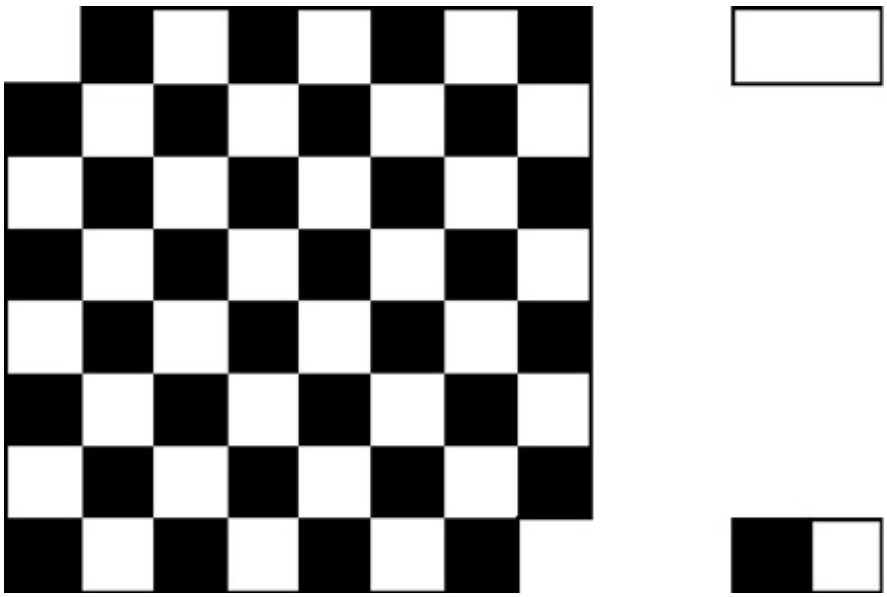
Poincaré zadał sobie wiele trudu, żeby podkreślić, iż bez początkowego etapu przygotowań uzyskanie jakiegokolwiek postępu prac jest mało prawdopodobne. Twierdził, że podświadomość potrzebuje dużej ilości materiału do rozmyślań, by mogła powstać szczęśliwa kombinacja idei prowadząca do ostatecznego rozwiązania. Bez pracy nie ma kołaczy. Musiał również wiedzieć – ponieważ wie to każdy twórczy matematyk – że ten prosty trójetapowy proces rzadko kiedy zachodzi tylko raz. Rozwiązanie problemu wymaga często dokonania kilku przełomów. Etap dojrzewania jednej idei może zostać przerwany przez dodatkowy proces przygotowania, dojrzewania i olśnienia czegoś, co jest potrzebne do dokończenia pierwotnego pomysłu. Rozwiązanie każdego problemu wartego zachodu, bez względu na to, czy jest wielki, czy nie, wymaga zwykle przeprowadzenia kilku takich procesów, zagnieżdżonych w sobie niczym zawiłe fraktale Benoîta Mandelbrota. Aby rozwiązać problem, należy go rozbić na podproblemy. Najpierw trzeba się jednak upewnić, że po ich rozwiązaniu będziemy potrafili złożyć wyniki

w całość i uzyskać rozwiązanie wyjściowego problemu. Jeśli tak, to możemy zająć się podproblemami. Czasami udaje się któryś rozwiązać, a czasami nie, i wtedy wypada wszystko na nowo przemyśleć. W niektórych wypadkach podproblem sam rozpada się na więcej części. Samo zarządzanie przebiegiem takich prac może być nie lada wyzwaniem.

Mówiąc o działaniu podświadomości, użyłem określenia „intuicja". Jest to jedno z tych uwodzicielskich słów, takich jak „instynkt", których powszechnie się używa, mimo że są całkowicie pozbawione jakiegokolwiek rzeczywistego znaczenia. Jest to nazwa czegoś, co podświadomie wyczuwamy, choć nie potrafimy tego zrozumieć. Intuicja matematyczna jest zdolnością umysłu do wykrywania kształtu, struktury i wzorców, których nie możemy dostrzec w sposób świadomy. Intuicja nie ma kryształowej przejrzystości świadomego wywodu logicznego, ale nadrabia ten brak, przyciągając naszą uwagę do kwestii, których nigdy świadomie byśmy nie rozważali. Neurobiolodzy dopiero zaczynają rozumieć, jak mózg realizuje znacznie prostsze zadania. Jednak bez względu na to, jak działa intuicja, musi ona wynikać ze struktury mózgu i tego, jak oddziałuje on ze światem zewnętrznym.

Często kluczowy wkład intuicji polega na uświadomieniu nam słabych punktów danego problemu, miejsc, w których można by go zaatakować. Dowód matematyczny przypomina bitwę lub też, jeśli wolicie mniej wojskowe porównanie, partię szachów. Po wykryciu potencjalnego słabego punktu badacz może wytoczyć wszystkie matematyczne działa, którymi potrafi się posługiwać, i spróbować wykorzystać zauważoną przewagę. Podobnie jak Archimedes prosił o wskazanie mocnego punktu podparcia, żeby poruszyć Ziemię, tak matematyk musi znaleźć jakiś sposób, by móc wykorzystać swoje umiejętności w zmaganiach z danym problemem. Wystarczy jeden dobry pomysł, żeby znaleźć punkt zaczepienia i zaatakować problem za pomocą standardowych metod. Reszta to już tylko umiejętne żonglowanie odpowiednimi technikami.

Moim ulubionym przykładem pokazującym, jak można znaleźć taki punkt oparcia, jest pewna zagadka, która nie ma głębszego matematycznego znaczenia, ale pozwala uświadomić sobie pewną ważną kwestię. Wyobraźmy sobie, że mamy standardową szachownicę o 64 kwadratowych polach i zapas kostek domina o rozmiarze odpowiednim do zakrycia dwóch sąsiednich pól szachownicy. Zakrycie całej szachownicy 32 kostkami domina jest dziecinnie łatwe. Załóżmy jednak, że z szachownicy usunięto dwa przeciwległe rogi, tak jak to pokazano na rycinie 1. Czy pozostałe 62 pola można przykryć za pomocą 31 kostek domina? Jeśli spróbujemy tego dokonać, szybko okaże się, że w żaden sposób nie można tego zrobić, chociaż trudno wskazać jakąś oczywistą przyczynę, która by to uniemożliwiała. Przełom nastąpi dopiero wtedy, gdy uświadomimy sobie, że jakkolwiek położymy kostkę domina, to zawsze będzie ona przykrywała jedno białe i jedno czarne pole. To jest nasz punkt podparcia. Teraz trzeba tylko go wykorzystać. Wynika z tego, że dowolny obszar szachownicy zakryty kostkami domina musi zawierać taką samą liczbę pól białych i czarnych. Jednak przeciwległe narożniki szachownicy są tego samego koloru, ich usunięcie więc (w naszym przykładzie są to białe pola) prowadzi do uzyskania figury zawierającej więcej pól czarnych niż białych – dokładnie rzecz biorąc, liczba pól czarnych jest o dwa większa od liczby białych. Takiej figury nie można zatem przykryć kostkami domina. Punktem podparcia w rozwiązaniu tej zagadki okazało się zwrócenie uwagi na to, jaka jest kombinacja kolorów pól przykrywanych przez *każdą* kostkę domina. Tym sposobem znaleźliśmy miejsce, w którym możemy zakotwiczyć rozumowanie logiczne i poprowadzić je dalej do końca. Gdybyśmy byli średniowiecznymi rycerzami szturmującymi zamek, powiedzielibyśmy, że musimy wykryć słabe miejsce w jego murach obronnych – punkt, na którym należy skoncentrować siłę ogniową katapult lub wykopać pod nim tunel, by w ten sposób go osłabić.



Ryc. 1. Czy taką wyszczerbioną szachownicę można pokryć kostkami domina, które przykrywają dokładnie dwa jej pola (u góry po prawej)? Jeśli odpowiednio pokolorujemy kostkę domina (u dołu po prawej) i policzymy, ile jest białych i czarnych pól, odpowiedź będzie oczywista.

Badania matematyczne różnią się od działań wojennych pod jednym ważnym względem. Każdy obszar raz zdobyty pozostaje już nasz na zawsze. Możemy bez obaw przerzucić wszystkie siły w inne miejsce, ponieważ raz udowodnione twierdzenie nie zniknie. Dzięki temu matematycy mogą dokonywać kolejnych postępów w pracy nad jakimś problemem, nawet jeśli nie uda im się go rozwiązać do końca. Wystarczy, że ustalą jakiś nowy fakt – ich odkrycie stanie się dostępne dla wszystkich i inni matematycy mogą je wykorzystać w dowolnym kontekście. Bardzo często punktem wyjścia do przypuszczenia nowego ataku na uporczywy problem jest dostrzeżenie jakiegoś matematycznego klejnotu, który niezauważony przez nikogo leżał zakopany w bezkształtnej masie nagromadzonych faktów. To jeden z powodów, dla których odkrywanie nowych obszarów matematyki jest ważne, nawet jeśli trudno znaleźć dla nich od razu jakieś zastosowanie. Tym sposobem powiększamy nasze terytorium, wzbogacamy zbrojownię o jeszcze jeden rodzaj broni.

Kiedyś może nadejdzie taki czas, że ten nowy obszar okaże się bezcenny – a równocześnie możemy być pewni, że to nie nastąpi, jeśli z góry założymy, iż jest on „bezużyteczny", i zapomnimy o nim lub nigdy go nie odkryjemy, tylko dlatego, że nikt nie znalazł dla niego żadnego *zastosowania*.

2 Simon Singh, *Tajemnica Fermata*, przeł. Paweł Strzelecki, Prószyński i S-ka, Warszawa 1999.

3 Gauss w liście do Heinricha Olbersa z 21 marca 1816 roku.

4 Tytuł ten brzmiał: *Krzywe modularne, formy eliptyczne i reprezentacje Galois*.

5 Andrew Wiles, *Modular elliptic curves and Fermat's last theorem* (*Modularne krzywe eliptyczne a wielkie twierdzenie Fermata*), „Annals of Mathematics" 1995, tom 141, s. 443–551.

6 Ian Stewart, *17 równań, które zmieniły świat*, przeł. Julia Szajkowska, Prószyński i S-ka, Warszawa 2013, rozdział 11.

7 *Ibid.*, rozdział 9.

8 Problemy Hilberta przedstawiłem w książce *Gabinet zagadek matematycznych*. Ich uaktualniona lista wygląda następująco:

1. **Hipoteza continuum:** Czy istnieje nieskończona liczba kardynalna plasująca się między mocą zbioru liczb całkowitych a mocą zbioru liczb rzeczywistych? Problem rozwiązany przez Paula Cohena w 1963 roku – odpowiedź zależy od wyboru aksjomatów teorii zbiorów.
2. **Spójność logiczna arytmetyki:** Należy udowodnić, że standardowe aksjomaty arytmetyki nigdy nie prowadzą do sprzeczności. Problem rozwiązał Kurt Gödel w 1931 roku – jest to niemożliwe w wypadku stosowanych zazwyczaj aksjomatów.
3. **Równość objętości czworościanów:** Jeśli mamy dane dwa czworościany o takiej samej objętości, to czy zawsze jest możliwe podzielenie jednego z nich na skończoną liczbę mniejszych wielościanów w taki sposób, by można je było złożyć w całość i otrzymać drugi czworościan? Problem rozwiązany w 1901 roku przez Maxa Dehna – okazuje się, że nie zawsze jest to możliwe.

4. **Prosta jako najkrótsza droga między dwoma punktami:** Należy sformułować aksjomaty geometrii w oparciu o powyższą definicję „prostej" i zbadać, do czego to prowadzi. Jest to zbyt szeroki problem, by można było przedstawić konkretne rozwiązanie, ale wykonano wiele prac poświęconych temu zagadnieniu.

5. **Grupy Liego bez założenia różniczkowalności:** Szczegółowy problem z zakresu teorii grup przekształceń. Dla jednej z interpretacji problem ten rozwiązał Andrew Gleason w latach pięćdziesiątych. Dla innej – rozwiązanie przedstawił Hidehiko Yamabe.

6. **Aksjomaty fizyki:** Należy opracować ścisły układ aksjomatów matematycznych obszarów fizyki, takich jak teoria prawdopodobieństwa i mechanika. Andriej Kołmogorow przedstawił aksjomaty teorii prawdopodobieństwa w 1933 roku.

7. **Liczby niewymierne i przestępne:** Należy udowodnić, że określone liczby są niewymierne lub przestępne. Problem rozwiązali Aleksander Gelfond i Theodor Schneider w 1934 roku.

8. **Hipoteza Riemanna:** Należy udowodnić, że wszystkie nietrywialne miejsca zerowe funkcji dzeta Riemanna leżą na prostej krytycznej. Zob. rozdz. 9.

9. **Prawo wzajemności ciał liczbowych:** Należy uogólnić klasyczne prawo wzajemności reszt kwadratowych na wyższe potęgi. Problem częściowo rozwiązany.

10. **Ustalenie, kiedy równanie diofantyczne ma rozwiązania:** Należy znaleźć algorytm, który pozwoli stwierdzić, kiedy równanie wielomianowe z wieloma zmiennymi ma rozwiązania w zbiorze liczb naturalnych. W 1970 roku Jurij Matijasewicz udowodnił, że jest to niemożliwe.

11. **Formy kwadratowe ze współczynnikami będącymi liczbami algebraicznymi:** Kwestia szczegółowa dotycząca rozwiązań równań diofantycznych o wielu zmiennych. Częściowo rozwiązany.

12. **Twierdzenie Kroneckera o ciałach abelowych:** Problem szczegółowy dotyczący uogólnienia twierdzenia Kroneckera. Wciąż nierozwiązany.

13. **Rozwiązanie równań stopnia siódmego za pomocą funkcji specjalnych:** Należy udowodnić, że w ogólnym przypadku nie da się rozwiązać równania stopnia siódmego za pomocą funkcji dwóch zmiennych. Dla jednej z interpretacji Andriej Kołmogorow i Władimir Arnold udowodnili twierdzenie przeciwne głoszące, że jest to możliwe.

14. **Skończoność pewnej struktury funkcji:** Należy rozszerzyć twierdzenie Hilberta o niezmiennikach algebraicznych na wszystkie grupy przekształceń. W 1959 roku Masayoshi Nagata udowodnił, że jest to niemożliwe.

15. **Rachunek Schuberta:** Hermann Schubert zaproponował nieścisłą metodę zliczania różnych konfiguracji geometrycznych, należy przedstawić ścisłą wersję tego rachunku. Jak dotąd nie znaleziono pełnego rozwiązania.

16. **Topologia krzywych i powierzchni:** Ile połączonych ze sobą elementów może mieć krzywa algebraiczna danego stopnia? Ile różnych cykli okresowych

może mieć algebraiczne równanie różniczkowe danego stopnia? Przeprowadzono pewne badania tych zagadnień, ale nie są one zbyt zaawansowane.

17. **Wyrażenie ściśle określonych form za pomocą kwadratów:** Czy funkcja wymierna przyjmująca zawsze wartości nieujemne musi być sumą kwadratów? Problem rozwiązali Emil Artin, D.W. Dubois i Albrecht Pfister – jest to prawdą dla liczb rzeczywistych, natomiast dla pozostałych systemów liczbowych stwierdzenie takie jest nieprawdziwe.

18. **Pokrycie przestrzeni wielościanami:** Ogólny problem wypełnienia przestrzeni wielokątami przystającymi. Problem ten dotyczy również hipotezy Keplera, która została udowodniona (zob. rozdz. 5).

19. **Analityczność rozwiązań rachunku wariacyjnego:** Rachunek wariacyjny odpowiada na pytania typu: „Jaka jest najkrótsza krzywa o danych własnościach?". Czy wystarczy zdefiniować taki problem za pomocą funkcji pozbawionych osobliwości, żeby rozwiązanie również ich nie miało? Udowodnił to w 1957 roku Ennio de Giorgi oraz, niezależnie od niego, John Nash.

20. **Problemy wartości brzegowej:** Należy opisać rozwiązania fizycznych równań różniczkowych we wnętrzu określonego obszaru przestrzeni, gdy dane są własności rozwiązań na brzegu tego obszaru. Problem został w zasadzie rozwiązany dzięki pracy wielu matematyków.

21. **Istnienie równań różniczkowych o danych grupach monodromii:** Specjalny rodzaj zespolonych równań różniczkowych można opisać za pomocą ich punktów osobliwych i grupy monodromii. Czy da się udowodnić, że może wystąpić dowolna kombinacja tych danych? Odpowiedź na tak postawione pytanie jest twierdząca lub przecząca, w zależności od przyjętej interpretacji.

22. **Uniformizacja relacji analitycznych za pomocą funkcji automorficznych:** Szczegółowy problem dotyczący uproszczenia równań. Rozwiązany przez Paula Koebego tuż po roku 1900.

23. **Rozwój rachunku wariacyjnego:** Hilbert apelował o przedstawienie nowych idei z dziedziny rachunku wariacyjnego. Zrobiono w tej dziedzinie już bardzo wiele, jest to jednak problem sformułowany zbyt ogólnie, żeby można go było uważać za rozwiązany.

2. Królestwo liczb pierwszych

Hipoteza Goldbacha

Niektóre wielkie problemy pojawiają się już na początku naszej edukacji matematycznej, choć najczęściej tego nie zauważamy. Zaraz po opanowaniu sztuki mnożenia natrafiamy na pojęcie liczb pierwszych. Zauważamy, że niektóre liczby można uzyskać przez wymnożenie ze sobą dwóch innych, mniejszych – na przykład: $6 = 2 \times 3$. Inne, takie jak 5, nie dają się rozbić w ten sposób – możemy najwyżej zauważyć, że $5 = 1 \times 5$, co jednak nie jest rozbiciem naszej liczby na dwie *mniejsze* wartości. Liczby dające się w ten sposób rozbić nazywamy liczbami złożonymi, natomiast te, których nie można tak podzielić, nazywamy liczbami pierwszymi. Liczby pierwsze wydają się tak proste. Aby je zrozumieć, wystarczy nauczyć się mnożenia. Są one podstawowymi składnikami liczb całkowitych i pojawiają się we wszystkich działach matematyki. Sprawiają również wrażenie niezwykle tajemniczych i wydaje się, że są rozmieszczone w sposób przypadkowy. Liczby pierwsze są całkowitą zagadką – nie ma co do tego wątpliwości. Być może wynika to z ich definicji, która zamiast mówić o tym, czym one są, opisuje czym nie są. Równocześnie odgrywają w matematyce podstawową rolę, nie możemy więc po prostu załamać bezradnie rąk i się poddać. Musimy dobrze je zrozumieć i odkryć ich najgłębsze sekrety.

Niektóre cechy są oczywiste. Z wyjątkiem najmniejszej liczby pierwszej, czyli liczby 2, wszystkie pozostałe są nieparzyste. Z wyjątkiem liczby 3 suma ich cyfr nie może być podzielna przez 3. Z wyjątkiem liczby 5 ostatnią cyfrą liczby pierwszej nie może być 5. Poza tymi przypadkami i kilkoma bardziej złożonymi regułami nie

znamy żadnych metod, które pozwalałyby z miejsca stwierdzić, czy dana liczba jest liczbą pierwszą. Istnieją wprawdzie wzory na liczby pierwsze, ale w większości wypadków są one oszustwem: nie dostarczają żadnej użytecznej nowej informacji na ich temat. Stanowią jedynie sprytne sposoby na zapisanie definicji liczby pierwszej w postaci wzoru. Liczby pierwsze są jak ludzie: są indywidualistkami i nie przestrzegają standardowych reguł.

W ciągu minionych tysiącleci matematycy powoli poszerzali naszą wiedzę na temat liczb pierwszych i od czasu do czasu udawało im się rozwiązać jakiś kolejny wielki problem z nimi związany. Jednak wciąż jeszcze wiele pytań pozostaje bez odpowiedzi. Niektóre mają charakter podstawowy i można je łatwo sformułować, inne dotyczą bardziej zawiłych kwestii. W tym rozdziale powiemy o tym, co wiemy i czego nie wiemy na temat tych denerwujących, ale podstawowych liczb. Na początku wyjaśnimy niektóre elementarne pojęcia, a w szczególności zajmiemy się rozkładem na czynniki pierwsze, czyli powiemy, jak można przedstawić daną liczbę za pomocą iloczynu liczb pierwszych. Nawet ta znana wszystkim procedura prowadzi na głęboką wodę, gdy tylko zaczniemy się zastanawiać nad naprawdę skutecznymi metodami znajdowania czynników pierwszych danej liczby. Zaskakujące jest na przykład to, że sprawdzenie, czy dana liczba jest liczbą pierwszą, czyli przeprowadzenie jej testu pierwszości, wydaje się względnie proste, ale jeśli okaże się liczbą złożoną, to znalezienie jej czynników pierwszych jest nierzadko zadaniem znacznie trudniejszym.

Po wyjaśnieniu podstawowych kwestii przejdziemy do najsłynniejszego nierozwiązanego problemu związanego z liczbami pierwszymi – do hipotezy Goldbacha, która już od 250 lat czeka na udowodnienie. W ostatnich latach dokonaliśmy ważnych postępów na drodze do osiągnięcia tego celu, ale samego dowodu nie udało się jeszcze przedstawić. Na koniec przedstawimy krótko kilka innych problemów, które pozwolą nam zrozumieć, co jeszcze czeka na odkrycie w tym bogatym, ale niesfornym obszarze matematyki.

Na lekcjach matematyki wszyscy poznajemy liczby pierwsze i rozkład na czynniki pierwsze, ale na tym poziomie edukacji rzadko kiedy wspomina się o najciekawszych cechach tych liczb, a jeśli już, to omawia się je bez dowodu. Są ku temu ważne powody: dowody nawet pozornie oczywistych własności są zaskakująco trudne. Zamiast tego uczniom pokazuje się proste działania na liczbach pierwszych i cały nacisk kładzie się na obliczenia z wykorzystaniem dość małych liczb. W efekcie nasze pierwsze zetknięcie z liczbami pierwszymi wprowadza nas do pewnego stopnia w błąd.

Starożytni Grecy znali niektóre podstawowe własności liczb pierwszych i wiedzieli, jak można udowodnić, że dana liczba jest liczbą pierwszą. Liczby pierwsze i czynniki pierwsze są głównym tematem VII księgi Euklidesowych *Elementów*, wielkiego klasycznego podręcznika geometrii. W tej konkretnej księdze przedstawiono geometryczne ujęcie arytmetycznego dzielenia i mnożenia. Grecy woleli operować na długościach odcinków zamiast na samych liczbach, ale bez większego wysiłku można wyrazić uzyskane przez nich wyniki w języku liczb. Euklides zadaje sobie wiele trudu, żeby udowodnić stwierdzenia, które mogą wydawać się oczywiste. Na przykład w twierdzeniu 16 z księgi VII dowodzi, że gdy mnożymy przez siebie dwie liczby, to wynik takiego działania nie zależy od kolejności, w jakiej je pomnożymy. Innymi słowy: $ab = ba$ – jest to jedno z podstawowych praw algebry.

W szkole wykorzystuje się czynniki pierwsze do znalezienia największego wspólnego dzielnika dwóch liczb. Aby na przykład znaleźć największy wspólny dzielnik liczb 135 i 630, rozkładamy je na czynniki pierwsze:

$$135 = 3^3 \times 5$$

$$630 = 2 \times 3^2 \times 5 \times 7$$

Następnie dla każdej znalezionej liczby pierwszej bierzemy największą potęgę, jaka występuje w obu rozkładach, i w ten sposób uzyskujemy: $3^2 \times 5$. Po wymnożeniu mamy wartość 45 – to jest

największy wspólny dzielnik. Patrząc na tę metodę, można odnieść wrażenie, że do znalezienia największego wspólnego dzielnika konieczny jest rozkład na czynniki pierwsze. W rzeczywistości związek logiczny między obiema operacjami jest odwrotny. Twierdzenie 2 z VII księgi *Elementów* przedstawia metodę na znajdowanie największego wspólnego dzielnika dwóch liczb całkowitych bez konieczności dokonywania ich rozkładu na czynniki pierwsze. Jej działanie opiera się na wielokrotnym odejmowaniu mniejszej liczby od większej. Należy stosować tę samą procedurę do otrzymanej reszty i mniejszej z dwóch liczb wyjściowych tak długo, aż nie zostanie nam żadna reszta. Dla wartości 135 i 630, czyli typowego przykładu z wykorzystaniem niedużych liczb, cały proces przebiega następująco: najpierw odejmujemy wielokrotnie 135 od 630

$$630 - 135 = 495,$$
$$495 - 135 = 360,$$
$$360 - 135 = 225,$$
$$225 - 135 = 90.$$

Ponieważ 90 jest mniejsze od 135, zamieniamy obie liczby miejscami:

$$135 - 90 = 45$$

Ponieważ 45 jest mniejsze od 90, zamieniamy obie liczby miejscami:

$$90 - 45 = 45$$
$$45 - 45 = 0$$

Zatem największym wspólnym dzielnikiem liczb 135 i 630 jest 45.

Metoda ta działa dlatego, że w każdym kroku zastępujemy pierwotną parę liczb prostszą parą (jedna z liczb jest zawsze mniejsza) o takim samym wspólnym największym dzielniku. W końcu

jedna z tych liczb dzieli się bez reszty i wtedy działanie procedury dobiega końca. W dzisiejszych czasach jasno zdefiniowaną metodę obliczeniową, która gwarantuje rozwiązanie danego zadania, nazywamy „algorytmem”. Dlatego procedurę Euklidesa nazywamy obecnie algorytmem Euklidesa. Z punktu widzenia logiki poprzedza on rozkład na czynniki pierwsze. Euklides wykorzystuje nawet ten algorytm do udowodnienia podstawowych własności czynników pierwszych – tak samo zresztą postępują obecnie wykładowcy matematyki na uniwersytetach.

Kluczową rolę w tym wszystkim odgrywa Euklidesowe twierdzenie 30. Mówiąc współczesnym językiem, stwierdza ono, że jeśli jakaś liczba pierwsza jest dzielnikiem iloczynu dwóch innych liczb – czyli wyniku ich wymnożenia – to musi być dzielnikiem jednej z nich. Twierdzenie 32 głosi, że każda liczba jest albo liczbą pierwszą, albo ma jakiś czynnik pierwszy. Po złożeniu tych dwóch twierdzeń razem nietrudno wywnioskować, że każda liczba jest iloczynem czynników pierwszych i że taki jej zapis jest jednoznaczny – może się jedynie zmieniać kolejność wystąpienia poszczególnych czynników pierwszych w iloczynie. Na przykład:

$$60 = 2 \times 2 \times 3 \times 5 = 2 \times 3 \times 2 \times 5 = 5 \times 3 \times 2 \times 2$$

i tak dalej, ale jedynym sposobem na uzyskanie liczby 60 jest przetasowanie tych czynników pierwszych. Nie jest na przykład możliwe przedstawienie rozkładu tej wartości na czynniki pierwsze postaci: $60 = 7 \times$ *coś*. Samo istnienie rozkładu na czynniki pierwsze wynika z twierdzenia 32. Jeśli dana liczba jest liczbą pierwszą, zakończ algorytm. Jeśli nie, znajdź jakiś czynnik pierwszy tej liczby, podziel ją przez znalezioną liczbę, by uzyskać mniejszą wartość, i powtórz algorytm od początku. Jednoznaczność wynika natomiast z twierdzenia 30. Gdyby na przykład istniał rozkład postaci $60 = 7 \times$ *coś*, to 7 musiałoby być dzielnikiem liczby 2, 3 lub 5, a tak nie jest.

W tym miejscu musimy sobie wyjaśnić pewną nieskomplikowaną, ale ważną kwestię: wyjątkowe znaczenie liczby 1. Zgodnie z przedstawioną wcześniej definicją, 1 jest bez wątpienia liczbą

pierwszą: jeśli spróbujemy rozłożyć liczbę 1 na czynniki pierwsze, to możemy najwyżej stwierdzić, że $1 = 1 \times 1$, a w tym iloczynie nie występują liczby mniejsze od naszej wartości wyjściowej. Jednak taka interpretacja prowadzi do pojawienia się problemów w bardziej zaawansowanych rozważaniach, zatem od mniej więcej dwóch stuleci matematycy dodają jeszcze jedno ograniczenie. Liczba 1 jest tak szczególna, że nie można jej uważać ani za liczbę pierwszą, ani za złożoną. Jest ona stworzeniem trzeciego rodzaju – wartością jednostkową. Jednym z powodów, dla których traktuje się 1 jako przypadek szczególny, a nie prawdziwą liczbę pierwszą, jest to, że jeśli nazwiemy 1 liczbą pierwszą, to jednoznaczność rozkładu przestaje obowiązywać. Wadę tę widać już w zapisie $1 \times 1 = 1$, a zapis $1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 = 1$ jest już kpiną w żywe oczy. Moglibyśmy zmodyfikować definicję jednoznaczności, mówiąc „rozkład jest jednoznaczny z wyjątkiem dodatkowych czynników o wartości 1”, ale byłby to jedynie inny sposób przyznania, że 1 jest liczbą szczególną.

Znacznie później, w twierdzeniu 20 z księgi IX, Euklides dowodzi prawdziwości kolejnego kluczowego faktu: „Liczb pierwszych jest więcej niż dowolna zadana ich liczba”. Innymi słowy, liczb pierwszych jest nieskończenie wiele. To cudowne twierdzenie ze sprytnym dowodem, ale otwiera ono prawdziwą puszkę Pandory. Jeśli liczby pierwsze ciągną się w nieskończoność, a przy tym pojawiają się bez żadnej prawidłowości, to jak możemy opisać, jak wyglądają?

Musimy się zmierzyć z tym pytaniem, ponieważ nie możemy zignorować liczb pierwszych. Są one podstawowym elementem matematycznego krajobrazu. Szczególnie często pojawiają się w teorii liczb, w której spełniają użyteczną funkcję. Ten dział matematyki zajmuje się badaniem własności liczb naturalnych. Może się wydawać, że to dosyć podstawowa kwestia, ale w istocie teoria liczb jest jedną z najgłębszych i najtrudniejszych gałęzi matematyki. W dalszej części książki sami się przekonamy o prawdziwości tego stwierdzenia. W 1801 roku Carl Friedrich Gauss, czołowy teoretyk liczb swojej epoki – i jeden z największych matematyków wszech

czasów, a może nawet największy – napisał zaawansowany podręcznik teorii liczb zatytułowany *Disquisitiones Arithmeticae* (*Badania arytmetyczne*). Przy okazji omawiania bardziej skomplikowanych zagadnień zamieścił tam uwagę, że nie powinniśmy tracić z oczu dwóch bardzo podstawowych kwestii: „Jak wiadomo, metody odróżniania liczb pierwszych od złożonych i rozkładania tych drugich na czynniki pierwsze są jednymi z najważniejszych i najbardziej użytecznych narzędzi stosowanych w arytmetyce".

W szkole poznajemy zazwyczaj jedną metodę znajdowania czynników pierwszych danej liczby: wypróbuj po kolei wszystkie możliwe czynniki, aż znajdziesz taki, który dzieli tę liczbę bez reszty. Jeśli dotrzesz do pierwiastka kwadratowego badanej liczby – a mówiąc dokładniej, do największej liczby naturalnej, która jest mniejsza od tego pierwiastka lub mu równa – i mimo to nie uda ci się wykryć żadnego czynnika pierwszego, to liczba ta jest liczbą pierwszą. W przeciwnym wypadku, gdy znajdziesz czynnik pierwszy, podziel przez niego badaną liczbę i powtórz algorytm od początku. Metoda ta jest najefektywniejsza, gdy bierze się pod uwagę tylko kolejne liczby pierwsze, ale w tym celu trzeba mieć ich listę. Poszukiwań można zaprzestać po dotarciu do pierwiastka kwadratowego badanej liczby, ponieważ najmniejszy czynnik pierwszy dowolnej liczby złożonej jest nie większy od jej pierwiastka kwadratowego. Procedura ta staje się jednak beznadziejnie nieskuteczna, gdy mamy do czynienia z naprawdę dużymi liczbami. Jeśli na przykład badaną liczbą jest:

$$1\,080\,813\,321\,843\,836\,712\,253,$$

której rozkład na czynniki pierwsze wygląda następująco:

$$13\,929\,010\,429 \times 77\,594\,408\,257,$$

to musielibyśmy wypróbować najpierw 624 401 249 kolejnych liczb pierwszych, zanim dotarlibyśmy do mniejszego z tych dwóch

czynników. Oczywiście, jeśli mamy do dyspozycji komputer, to zadanie takie nie jest zbyt trudne, ale jeśli weźmiemy jakąś stucyfrową liczbę, która jest iloczynem dwóch pięćdziesięciocyfrowych liczb pierwszych, i zastosujemy metodę systematycznego wypróbowywania kolejnych liczb pierwszych, to Wszechświat się skończy, zanim komputer znajdzie odpowiedź.

Prawda jest jednak taka, że komputery potrafią zwykle rozkładać na czynniki pierwsze liczby stucyfrowe. Mój komputer potrzebuje niecałej sekundy, żeby znaleźć czynniki pierwsze liczby $10^{99} + 1$, która wygląda jak 1000…001 z 98 zerami. Jest ona iloczynem 13 liczb pierwszych (jedna z nich występuje dwukrotnie), z których najmniejszą jest 7, a największą:

141 122 524 877 886 182 282 233 539 317 796 144 938
305 111 168 717

Jeśli jednak każę komputerowi znaleźć rozkład na czynniki pierwsze liczby $10^{199} + 1$, która ma 200 cyfr, to będzie liczył przez całe stulecia i nic nie osiągnie. Mimo to znalezienie rozkładu liczby stucyfrowej robi wrażenie. Na czym polega sekret? Musimy mieć jakąś lepszą metodę od wypróbowywania kolejnych liczb pierwszych.

Obecnie wiemy znacznie więcej niż Gauss na temat pierwszej ze wspomnianych przez niego metod (sprawdzania liczb pierwszych) i znacznie mniej, niżbyśmy chcieli, na temat drugiej (rozkładu na czynniki pierwsze). Uważa się powszechnie, że sprawdzenie, czy dana liczba jest liczbą pierwszą, jest dużo łatwiejsze od znalezienia jej rozkładu na czynniki pierwsze. Stwierdzenie takie jest zazwyczaj dużym zaskoczeniem dla osób niezajmujących się matematyką, które dowiedziały się w szkole, że sprawdzenie, czy dana liczba jest liczbą pierwszą, wymaga zastosowania tej samej metody jak w wypadku szukania rozkładu na czynniki pierwsze, a mianowicie należy wypróbować po kolei wszystkie możliwe dzielniki. Okazuje się, że są pewne sprytne sposoby na udowodnienie, że badana liczba jest liczbą pierwszą, bez uruchamiania tej żmudnej procedury. Nadają się

one również do udowodnienia, że liczba jest liczbą złożoną, ale bez podawania jej czynników pierwszych. Wystarczy jedynie pokazać, że liczba ta nie przechodzi testu na liczbę pierwszą.

Wielkim pradziadkiem wszystkich współczesnych metod wykrywania liczb pierwszych jest twierdzenie Fermata, które dla odróżnienia od słynnego wielkiego twierdzenia Fermata (omówimy je w rozdziale 7) nazywa się małym. Twierdzenie to opiera się na arytmetyce modularnej, którą czasami nazywa się również „arytmetyką zegarową", ponieważ liczby zawijają się w niej niczym godziny na tarczy zegarowej. Wybierzmy dowolną liczbę – w przypadku dwunastogodzinnego zegara analogowego jest to 12 – i nazwijmy ją modułem. Wykonując dowolne operacje na liczbach całkowitych, pozwalamy sobie teraz zastąpić dowolną wielokrotność wybranej liczby – czyli 12 – wartością zero. Na przykład $5 \times 5 = 25$, ale 24 jest dwukrotnością 12, a zatem po odjęciu 24 otrzymujemy $5 \times 5 = 1$ przy module wynoszącym 12. Arytmetyka modularna jest bardzo elegancka, ponieważ działają w niej niemal wszystkie tradycyjne reguły arytmetyki. Główna różnica polega na tym, że nie zawsze daje się podzielić przez siebie dwie liczby, nawet jeśli żadna z nich nie jest zerem. Arytmetyka modularna jest również bardzo przydatna, ponieważ pozwala nam rozprawić się w elegancki sposób z różnymi problemami związanymi z podzielnością: które liczby są podzielne przez wybrany moduł? A ile wynosi reszta z dzielenia, jeśli nie są podzielne? Gauss wprowadził arytmetykę modularną w *Disquisitiones Arithmeticae*, a obecnie stosuje się ją powszechnie w informatyce, fizyce i inżynierii, a także oczywiście w matematyce.

Małe twierdzenie Fermata głosi, że jeśli wybierzemy moduł p będący liczbą pierwszą, a następnie weźmiemy wartość a, która nie jest wielokrotnością p, to wtedy a do potęgi $(p - 1)$ jest równe 1 w arytmetyce z modułem p. Przyjmijmy na przykład, że $p = 17$ i $a = 3$. W takim razie z twierdzenia Fermata wynika, że jeśli podzielimy 3^{16} przez 17, uzyskamy resztę 1. Sprawdźmy:

$$3^{16} = 43\,046\,721 = 2\,532\,160 \times 17 + 1.$$

Nikt przy zdrowych zmysłach nie będzie zapewne próbował przeprowadzać takich obliczeń dla, powiedzmy, stucyfrowych liczb pierwszych. Na szczęście istnieje sprytny, szybki sposób na wykonanie tego rodzaju rachunków. Chodzi o to, że jeśli wynik nie jest równy 1, to oznacza to, że wybrana wartość modułu jest liczbą złożoną. Zatem małe twierdzenie Fermata pozwala na opracowanie skutecznego testu pierwszości i stanowi warunek konieczny tego, by dana liczba była liczbą pierwszą.

Niestety, nie jest to warunek wystarczający. Wiele liczb złożonych – są to tak zwane liczby Carmichaela – przechodzi taki test z pozytywnym wynikiem. Najmniejszą taką liczbą jest 561, a w 2003 roku Red Alford, Andrew Granville i Carl Pomerance udowodnili ku zdumieniu wszystkich, że jest ich nieskończenie wiele. Zdumienie wywołał fakt, że matematycy ci znaleźli taki dowód – sam wynik był mniej zaskakujący. Mówiąc ściśle, uczeni ci pokazali, że istnieje przynajmniej $x^{2/7}$ liczb Carmichaela mniejszych lub równych wartości x, jeśli x jest odpowiednio duże.

Bardziej wyrafinowane odmiany małego twierdzenia Fermata można jednak przekształcić w prawdziwe testy pierwszości. Jeden z takich sposobów przedstawił w 1976 roku Gary Miller. Niestety, dowód poprawności metody Millera zależy od jednego z nierozwiązanych wielkich problemów, a mianowicie od poprawności uogólnionej hipotezy Riemanna (powiemy o niej w rozdziale 9). W 1980 roku Michael Rabin przekształcił metodę Millera w procedurę probabilistyczną – jednak taki test pierwszości może od czasu do czasu dawać błędny wynik. Wyjątki, jeśli istnieją, są bardzo rzadkie, ale nie można ich całkowicie wykluczyć. Najskuteczniejszym jak dotąd deterministycznym testem pierwszości jest tak zwany test APR, nazwany tak od pierwszych liter nazwisk Leonarda Adlemana, Carla Pomerance'a i Roberta Rumely'ego. Wykorzystuje on znacznie bardziej skomplikowane pojęcia z zakresu teorii liczb niż małe twierdzenie Fermata, ale działa w podobny sposób.

Wciąż żywo pamiętam list, który otrzymałem od rozentuzjazmowanego matematyka amatora. Jego autor

zaproponował modyfikację klasycznego algorytmu dzielenia przez kolejne liczby pierwsze. Ta metoda również polegała na wypróbowywaniu kolejnych dzielników, ale algorytm zaczyna się od pierwiastka kwadratowego danej liczby i przesuwa *w dół*. Takie podejście czasami szybciej prowadzi do uzyskania wyniku niż wykonywanie tej procedury w tradycyjny sposób, ale przy dużych liczbach natrafia na takie same problemy jak algorytm standardowy. Jeśli wypróbujemy to podejście na przytoczonym wcześniej przykładzie z 22-cyfrową liczbą 1 080 813 321 843 836 712 253, to musimy wyjść od wartości bliskiej jej pierwiastka kwadratowego, czyli 32 875 725 419. W takim wypadku musimy wypróbować 794 582 971 liczb pierwszych, zanim dotrzemy do wartości, która podzieli naszą liczbę bez reszty. To jeszcze *gorszy* wynik niż w sytuacji poszukiwania dzielnika w tradycyjny sposób.

W 1956 roku słynny logik Kurt Gödel w liście do Johna von Neumanna powtórzył apel Gaussa. Zastanawiał się w nim, czy algorytm polegający na próbie dzielenia przez kolejne czynniki można by jakoś usprawnić, a jeśli tak, to jak bardzo. Von Neumann nie zainteresował się tym problemem, ale w następnych latach inni uczeni udzielili Gödlowi odpowiedzi, odkrywając praktyczne metody znajdowania liczb pierwszych, działające dla dużych, stucyfrowych liczb, a czasem i większych. Metody te, z których najpopularniejsza nazywa się sito kwadratowe, są znane od około roku 1980. Jednak niestety prawie wszystkie są albo probabilistyczne, albo okazują się nieefektywne w pewnym sensie, który zaraz wyjaśnię.

Jak rośnie czas działania algorytmu w miarę wzrostu rozmiaru danych wejściowych? W przypadku testu pierwszości rozmiarem danej wejściowej nie jest badana wartość, ale liczba jej cyfr. Podstawowe rozróżnienie podczas prowadzenia takich rozważań przebiega między dwiema klasami algorytmów, które oznacza się symbolami P i nie-P. Jeśli czas działania rośnie proporcjonalnie do stałej potęgi rozmiaru danej wejściowej, to taki algorytm zaliczamy do klasy P – w przeciwnym razie należy do klasy nie-P. Mówiąc ogólnie, algorytmy klasy P są użyteczne, natomiast algorytmy klasy nie-P są niepraktyczne, jednak między tymi skrajnościami istnieje pewien pas ziemi niczyjej, w którym znaczenia nabierają inne

czynniki. Symbol P pochodzi od angielskiego określenia *polynomial time* (czas wielomianowy), czyli nazwa ta odwołuje się w wyszukany sposób do potęg. (Do zagadnienia złożoności algorytmów powrócimy w rozdziale 11).

Patrząc z perspektywy algorytmów klasy P, należy stwierdzić, że procedura poszukiwania czynników pierwszych metodą dzielenia przez kolejne liczby spisuje się bardzo kiepsko. Nie odczuwa się tego być może w szkolnych zadaniach, ponieważ wówczas pojawiają się liczby o długości dwóch, a najwyżej trzech cyfr, ale dla liczb stucyfrowych metoda ta jest całkowicie beznadziejna. Nie ma wątpliwości, że należy do klasy nie-P. W istocie czas działania tego algorytmu dla *n*-cyfrowej liczby jest proporcjonalny mniej więcej do $10^{n/2}$, czyli rośnie znacznie szybciej niż dowolna stała potęga *n*. Ten rodzaj wzrostu, nazywany wykładniczym, jest *naprawdę* zły – to prawdziwe obliczeniowe grzęzawisko.

Do lat osiemdziesiątych XX wieku wszystkie znane algorytmy na test pierwszości, z wyjątkiem algorytmów probabilistycznych i takich, których poprawności nie udowodniono, miały złożoność wykładniczą. Jednak w 1983 roku odkryto algorytm leżący w owym pasie ziemi niczyjej bardzo blisko obszaru algorytmów klasy P – chodzi o wspomniany już test APR. Henri Cohen i Hendrik Lenstra zaproponowali ulepszoną wersję tego testu o czasie działania proporcjonalnym do *n* do potęgi ln ln *n*, gdzie „ln” oznacza logarytm (naturalny). Ściśle rzecz biorąc, ln ln *n* może przyjąć dowolnie dużą wartość, zatem ten algorytm nie jest klasy P. Nie wyklucza to jednak możliwości praktycznego zastosowania takiego algorytmu: jeśli *n* jest googolpleksem, czyli jedynką, za którą stoi 10^{100} zer, to ln ln *n* wynosi około 230. Jak głosi stary dowcip: „Udowodniono wprawdzie, że ln ln *n* zmierza do nieskończoności, ale nikomu nie udało się tego zobaczyć na własne oczy”.

Pierwszy test pierwszości należący do klasy P odkrył w 2002 roku Manindra Agrawal wraz ze swoimi ówczesnymi studentami Neerajem Kayalem i Nitinem Saxeną. W uwagach zamieściłem ogólny opis tego algorytmu[9]. Uczeni ci dowiedli, że czas działania ich algorytmu w najgorszym wypadku jest proporcjonalny do n^{12}. Czas ten udało

się wkrótce skrócić do wartości $n^{7,5}$. Jednak mimo że ich algorytm należy do klasy P, a zatem grupy procedur uznawanych za „wydajne”, jego zalety widoczne są dopiero wtedy, gdy *n* staje się naprawdę duże. Algorytm ten powinien pokonać test APR, gdy liczba *cyfr* liczby *n* wynosi około 10^{1000}. Tak duża liczba nie zmieści się w pamięci żadnego komputera, a prawdę mówiąc, nie uda się jej nawet zmieścić w znanym nam Wszechświecie. Jednak skoro teraz już *wiemy*, że istnieje przynajmniej jeden test pierwszości klasy P, możemy zacząć szukać lepszego algorytmu. Lenstrze i Pomerance'owi udało się zmniejszyć wykładnik w oszacowaniu złożoności algorytmu z 7,5 do 6. Jeśli uda się udowodnić parę innych hipotez na temat liczb pierwszych, to wykładnik ten zmniejszy się do wartości 3, a to już zaczyna wyglądać całkiem dobrze.

Najbardziej intrygującym aspektem algorytmu AKS (Agrawala, Kayala i Saxeny) nie jest jednak osiągnięty wynik, ale zastosowana metoda. Jest ona mianowicie prosta – przynajmniej dla matematyków – i nowatorska. Opiera się na pewnej odmianie małego twierdzenia Fermata, ale zamiast działań na liczbach zespół Agrawala zastosował operacje na wielomianach. Wielomian jest kombinacją potęg zmiennej *x*, na przykład: $5x^3 + 4x - 1$. Można je dodawać, odejmować i mnożyć z zachowaniem normalnych praw algebry. W rozdziale 3 wyjaśnimy wielomiany bardziej szczegółowo.

To był naprawdę cudowny pomysł: rozszerzmy obszar, na którym prowadzi się rozważania, i przenieśmy nasz problem na zupełnie nowy teren. Jest to jedna z tych idei, które są tak proste, że trzeba być geniuszem, by na nie wpaść. Swój początek wzięła z artykułu z 1999 roku, który Agrawal napisał wspólnie z opiekunem swojej pracy doktorskiej, Somenathem Biswasem. W artykule tym uczeni przedstawili probabilistyczny test pierwszości operujący wielomianami, oparty na małym twierdzeniu Fermata. Agrawal był przekonany, że z tego testu można usunąć czynnik probabilistyczny. W 2001 roku jego studenci odkryli pewien kluczowy, ale dość skomplikowany fakt. Podążając tym tropem, zespół wypłynął na głębokie wody teorii liczb, ale ostatecznie udało się wszystko poprawnie wyprowadzić i została już tylko jedna przeszkoda:

konieczność udowodnienia istnienia liczby pierwszej *p* takiej, że *p* – 1, ma odpowiednio duży czynnik pierwszy. Uczeni popytali trochę wkoło, poszukali w Internecie i dotarli do twierdzenia, które Étienne Fouvry udowodnił w 1985 roku z wykorzystaniem zaawansowanych, skomplikowanych metod. Właśnie tego było im trzeba, by udowodnić, że ich algorytm działa, i w ten sposób ostatni kawałek układanki znalazł się na swoim miejscu.

W czasach gdy teoria liczb siedziała bezpiecznie w swojej niewielkiej wieży z kości słoniowej, nic z tego, o czym tu mówimy, nie miałoby większego znaczenia dla reszty świata. Jednak w ciągu ostatnich dwudziestu lat liczby pierwsze znalazły ważne zastosowanie w kryptologii, czyli nauce o szyfrowaniu. Szyfry odgrywają ważną rolę nie tylko w zastosowaniach wojskowych – przedsiębiorstwa również mają swoje tajemnice. Prawdę mówiąc, w epoce Internetu wszyscy używamy szyfrów – nie chcemy przecież, by przestępcy uzyskali dostęp do naszych kont bankowych, numerów kart kredytowych, czy nawet, w obliczu coraz częstszych przypadków kradzieży tożsamości, by poznali imię naszego kota. A przecież Internet jest tak wygodnym narzędziem do płacenia rachunków, odnawiania polis ubezpieczeniowych i rezerwowania hoteli, że musimy pogodzić się z pewnym ryzykiem, iż nasze wrażliwe, osobiste dane mogą wpaść w niepowołane ręce.

Producenci komputerów i twórcy serwisów internetowych starają się zmniejszyć to ryzyko, stosując różne techniki szyfrowania. Użycie komputerów całkowicie odmieniło zarówno kryptologię, jak i kryptoanalizę, czyli mroczną sztukę łamania szyfrów. Opracowano wiele nowych szyfrów, a jeden z najsłynniejszych, wymyślony przez Rona Rivesta, Adiego Shamira i Leonarda Adlemana w 1978 roku, wykorzystuje liczby pierwsze. Duże, o długości około stu cyfr. Szyfr RSA (Rivesta, Shamira i Adlemana) jest obecnie zaimplementowany w wielu komputerowych systemach operacyjnych, wbudowany w główne protokoły bezpiecznej komunikacji internetowej i powszechnie korzystają z niego rządy, przedsiębiorstwa i uniwersytety. Nie oznacza to, że każde nowe odkrycie dotyczące

liczb pierwszych ma znaczenie dla bezpieczeństwa naszych internetowych kont bankowych. Dodatkowy dreszczyk emocji pojawia się tylko podczas dokonywania odkryć związanych z wyznaczaniem liczb pierwszych. Dobrym tego przykładem jest test AKS. Z matematycznego punktu widzenia jest elegancki i ważny, ale nie ma żadnego bezpośredniego znaczenia praktycznego.

Test AKS stawia jednak w nowym, nieco niepokojącym świetle ogólne bezpieczeństwo szyfru RSA. Wciąż nie znamy algorytmu klasy P, który realizowałby drugą z wymienionych przez Gaussa metod, czyli dokonywał rozkładu liczby na czynniki pierwsze. Większość ekspertów uważa, że taki algorytm nie istnieje, ale obecnie specjaliści nie są już tego tak pewni jak kiedyś. Ponieważ mogą się gdzieś czaić nowe odkrycia, takie jak test AKS, wykorzystujące tak proste idee jak wielomianowa wersja małego twierdzenia Fermata, to należy przyjąć, że układy szyfrujące wykorzystujące rozkład na czynniki pierwsze wcale nie muszą być tak bezpieczne, jak sobie wyobrażamy. Na wszelki wypadek, na razie nie powinniśmy jeszcze ujawniać w Internecie, jak się wabi nasz kot.

Nawet podstawowe reguły matematyczne związane z liczbami pierwszymi szybko prowadzą do bardziej zaawansowanych pojęć. Aura tajemniczości potęguje się, gdy zaczynamy stawiać bardziej subtelne pytania. Euklides dowiódł, że liczby pierwsze ciągną się w nieskończoność, nie możemy więc sporządzić ich pełnej listy i zapomnieć o całej sprawie. Nie możemy też podać prostego, użytecznego wzoru algebraicznego na wyznaczanie kolejnych liczb pierwszych, takiego jak wzór x^2 opisujący kolejne kwadraty liczb. (Istnieją wprawdzie proste wzory, ale są one „oszustwem", ponieważ liczby pierwsze są w nich wbudowane w sposób niejawny i w istocie nie mówią nic nowego)[10]. Aby poznać naturę tych nieuchwytnych, nieregularnych liczb, przeprowadzamy badania, poszukujemy jakichś śladów struktury i próbujemy udowodnić, że zauważone prawidłowości obowiązują bez względu na wielkość liczb pierwszych. Możemy na przykład zapytać, jak wygląda rozkład liczb pierwszych wśród wszystkich liczb naturalnych. Z analizy tablic liczb pierwszych

płynie wniosek, że występują one coraz rzadziej, w miarę jak stają się coraz większe. W tabeli 1 pokazano, ile liczb pierwszych występuje w różnych przedziałach obejmujących 1000 kolejnych liczb naturalnych.

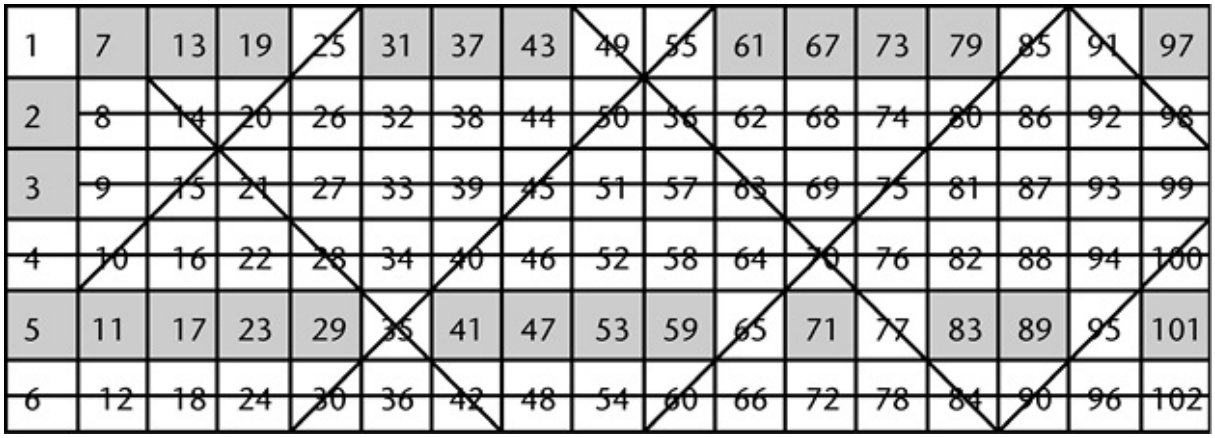
Przedział	**Ilość liczb pierwszych**
1–1000	168
1001–2000	135
2001–3000	127
3001–4000	119
4001–5000	118
5001–6000	114
6001–7000	117
7001–8000	106
8001–9000	110
9001–10 000	111

Tabela 1. Ilość liczb pierwszych w kolejnych przedziałach obejmujących 1000 liczb naturalnych

Liczby w drugiej kolumnie najczęściej maleją w miarę przesuwania się w dół tabeli, choć czasami występują krótkie okresy, gdy zaczynają rosnąć: po 114 pojawia się na przykład 117. Jest to przejaw nieregularności liczb pierwszych, ale mimo to występuje wyraźna ogólna tendencja zmniejszania się liczby liczb pierwszych, w miarę jak stają się one coraz większe. Wyjaśnienia tego faktu nie trzeba daleko szukać: im większa jest dana liczba, tym więcej istnieje jej potencjalnych czynników pierwszych. Liczby pierwsze muszą unikać wszystkich tych czynników. Przypomina to trochę odławianie liczb złożonych za pomocą sieci – im mniejsze będą oka sieci, tym mniej liczb pierwszych przez nie ucieknie.

Owa „sieć” ma nawet swoją nazwę: jest to sito Eratostenesa. Eratostenes z Cyreny był starożytnym greckim matematykiem, żyjącym około 250 roku p.n.e. Był również atletą, a poza tym interesował się poezją, geografią, astronomią i muzyką. Obserwując położenie Słońca w południe w dwóch różnych miejscowościach –

w Aleksandrii i Syene (dzisiejszy Asuan) – dokonał pierwszego wiarygodnego pomiaru rozmiaru Ziemi. W południe Słońce znajdowało się w Syene dokładnie nad jego głową, ale w Aleksandrii było wtedy widoczne pod kątem 7 stopni. Ponieważ kąt ten jest jedną pięćdziesiątą pełnego okręgu, Eratostenes doszedł do wniosku, że obwód Ziemi musi być 50 razy większy od odległości między Aleksandrią i Syene. Nie mógł zmierzyć bezpośrednio tak dużej odległości, zapytał więc kupców, jak długo trwa taka podróż na wielbłądzie, i oszacował, jaką drogę pokonuje średnio wielbłąd w ciągu dnia. Wynik swoich obliczeń wyraził w stadionach, ale niestety nie wiemy, jaką długość miała ta starożytna jednostka odległości. Historycy na ogół uważają, że oszacowanie Eratostenesa było dość dokładne.



Ryc. 2. Sito Eratostenesa

Sito Eratostenesa jest algorytmem wyznaczania wszystkich liczb pierwszych przez eliminowanie ich kolejnych wielokrotności. Na rycinie 2 pokazano zastosowanie tej metody dla liczb z przedziału od 1 do 102, ułożonych w taki sposób, żeby proces wykreślania wielokrotności można było łatwo prześledzić. Aby zrozumieć, co się wówczas dzieje, proponuję, żebyście sami skonstruowali taki diagram. Należy zacząć od samej siatki, pomijając linie wykreślające poszczególne liczby. Omijamy jedynkę, ponieważ jest wartością jednostkową. Następną liczbą jest 2, a zatem jest to liczba pierwsza.

Wykreślmy wszystkie wielokrotności 2: liczby te leżą w wierszach zaczynających się od 4, 6, i 8. Następną niewykreśloną liczbą jest 3, a zatem jest to liczba pierwsza. Wykreślmy wszystkie wielokrotności 3: liczby te znajdują się w wierszach zaczynających się od 6 (już wykreślony) i 9. Kolejną niewykreśloną liczbą jest 5, a więc jest to liczba pierwsza. Wykreślmy wszystkie wielokrotności 5: są to liczby leżące na przekątnych przebiegających w górę i w prawo, począwszy od liczby 10. Kolejną niewykreśloną liczbą jest 7, a zatem jest to liczba pierwsza. Wykreślmy wszystkie wielokrotności 7: są to liczby leżące na przekątnych przebiegających w dół i w prawo, począwszy od liczby 14. Następną niewykreśloną liczbą jest 11, a zatem jest to liczba pierwsza. Pierwszą wielokrotnością liczby 11, która nie została jeszcze wykreślona, jest liczba 121 – wszystkie poprzednie wielokrotności 11 zostały już wykreślone, ponieważ mają jeszcze mniejsze czynniki pierwsze. Liczba 121 wykracza już poza nasz diagram, a zatem możemy przerwać cały proces. Pozostałe niewykreślone liczby, zaznaczone szarym kolorem, są liczbami pierwszymi.

Sito Eratostenesa nie jest tylko historyczną ciekawostką – wciąż jest to jedna z najbardziej wydajnych metod sporządzania obszernych list liczb pierwszych. Metody bazujące na sicie Eratostenesa pozwoliły na uczynienie znacznych postępów na drodze do wyjaśnienia najsłynniejszego chyba nierozwiązanego wielkiego problemu dotyczącego liczb pierwszych: hipotezy Goldbacha. Niemiecki matematyk amator Christian Goldbach korespondował z wieloma słynnymi uczonymi swej epoki. W 1742 roku w liście do Leonharda Eulera zamieścił kilka ciekawych hipotez na temat liczb pierwszych. Później historycy odkryli, że mniej więcej to samo stwierdził kilka lat wcześniej Kartezjusz. Pierwsze stwierdzenie Goldbacha brzmi: „Każdą liczbę naturalną, którą można zapisać jako sumę dwóch liczb pierwszych, można również zapisać jako sumę dowolnej liczby liczb pierwszych, aż do sytuacji, gdy suma będzie składała się z samych jedności”. Drugie zdanie, dodane na marginesie, brzmi: „Każdą liczbę naturalną większą od 2 można przedstawić jako sumę trzech liczb

pierwszych”. Biorąc pod uwagę dzisiejszą definicję liczby pierwszej, należy zauważyć, że są pewne oczywiste odstępstwa od tych reguł. Na przykład 4 nie jest sumą trzech liczb pierwszych. Ponieważ najmniejszą liczbą pierwszą jest 2, to suma trzech liczb pierwszych musi wynosić przynajmniej 6. Natomiast w czasach Goldbacha 1 uznawano za liczbę pierwszą. Jego hipotezy można jednak bardzo łatwo sformułować tak, by uwzględniały współczesną definicję liczby pierwszej.

W swojej odpowiedzi Euler wspomniał ich wcześniejszą rozmowę, w której Goldbach zauważył, że jego pierwsza hipoteza wynika z prostszej tezy, jego trzeciej hipotezy: „Każda parzysta liczba naturalna jest sumą dwóch liczb pierwszych”. W sytuacji, gdy obowiązywała konwencja, że 1 jest liczbą pierwszą, z tego stwierdzenia wynika także jego druga hipoteza, ponieważ każdą liczbę można zapisać albo jako $n + 1$, albo jako $n + 2$, gdzie n jest parzyste. Jeśli n jest sumą dwóch liczb pierwszych, to badana liczba jest sumą trzech liczb pierwszych. Euler miał w odniesieniu do trzeciej hipotezy Goldbacha mieszane uczucia: „Uważam, że jest to całkowicie pewne twierdzenie, ale nie potrafię tego udowodnić”. Zdanie to dobrze oddaje również naszą obecną opinię w tej kwestii.

Współczesne rozumienie liczb pierwszych, zgodnie z którym 1 nie jest liczbą pierwszą, sprawia, że hipoteza ta rozpada się na dwa różne stwierdzenia. Hipoteza Goldbacha dla liczb parzystych brzmi:

Każda parzysta liczba naturalna większa od 2 jest sumą dwóch liczb pierwszych.

Natomiast hipoteza Goldbacha dla liczb nieparzystych ma postać:

Każda nieparzysta liczba naturalna większa od 5 jest sumą trzech liczb pierwszych.

Z hipotezy dla liczb parzystych wynika hipoteza dla liczb nieparzystych, ale nie odwrotnie[11]. Warto rozważać obie hipotezy oddzielnie, ponieważ wciąż nie wiemy, czy którakolwiek z nich jest

prawdziwa. Hipoteza dla liczb nieparzystych wydaje się nieco łatwiejsza, ponieważ udało się dokonać większych postępów na drodze do jej udowodnienia.

Wystarczy przeprowadzić kilka prostych obliczeń, żeby sprawdzić, że hipoteza Goldbacha dla liczb parzystych jest spełniona dla niedużych wartości:

$$4 = 2 + 2,$$
$$6 = 3 + 3,$$
$$8 = 5 + 3,$$
$$10 = 7 + 3 = 5 + 5,$$
$$12 = 7 + 5,$$
$$14 = 11 + 3 = 7 + 7,$$
$$16 = 13 + 3 = 11 + 5,$$
$$18 = 13 + 5 = 11 + 7,$$
$$20 = 17 + 3 = 13 + 7.$$

Bez trudu moglibyśmy ciągnąć te obliczenia do, powiedzmy, 1000 – a nawet dalej, jeśli starczyłoby nam cierpliwości. Na przykład 1000 = 3 + 997, a 1 000 000 = 17 + 999 993. W 1938 roku Nils Pipping sprawdził poprawność hipotezy Goldbacha dla wszystkich liczb parzystych aż do 100 000.

Jest również jasne, że w miarę jak rozważamy coraz większe liczby, mamy zazwyczaj do dyspozycji coraz więcej sposobów zapisania ich w postaci sumy liczb pierwszych. To całkiem sensowne. Wystarczy się zastanowić, co będzie, gdy weźmiemy jakąś dużą liczbę parzystą i zaczniemy od niej odejmować kolejne liczby pierwsze. Jakie będzie wówczas prawdopodobieństwo, że w wyniku *wszystkich* takich odejmowań uzyskamy liczby złożone? Do spełnienia hipotezy Goldbacha dla tej liczby wystarczy, żeby wśród powyższych wyników pojawiła się tylko jedna liczba pierwsza. Wykorzystując własności statystyczne liczb pierwszych, możemy oszacować prawdopodobieństwo uzyskania takiego wyniku. Analitycy Godfrey Harold Hardy i John Edensor Littlewood przeprowadzili takie obliczenia w 1923 roku i uzyskali wiarygodny, ale nieścisły wzór na

liczbę różnych sposobów wyrażenia danej liczby parzystej *n* jako sumy dwóch liczb pierwszych – takich sposobów jest w przybliżeniu $n/\left[2(\ln n)^2\right]$. Wartość ta rośnie ze wzrostem *n*, co zgadza się z zauważoną prawidłowością liczbową. Jednak nawet gdyby te obliczenia udało się uściślić, to i tak może wystąpić jakiś rzadki wyjątek, nie rozwiązuje to więc naszego problemu.

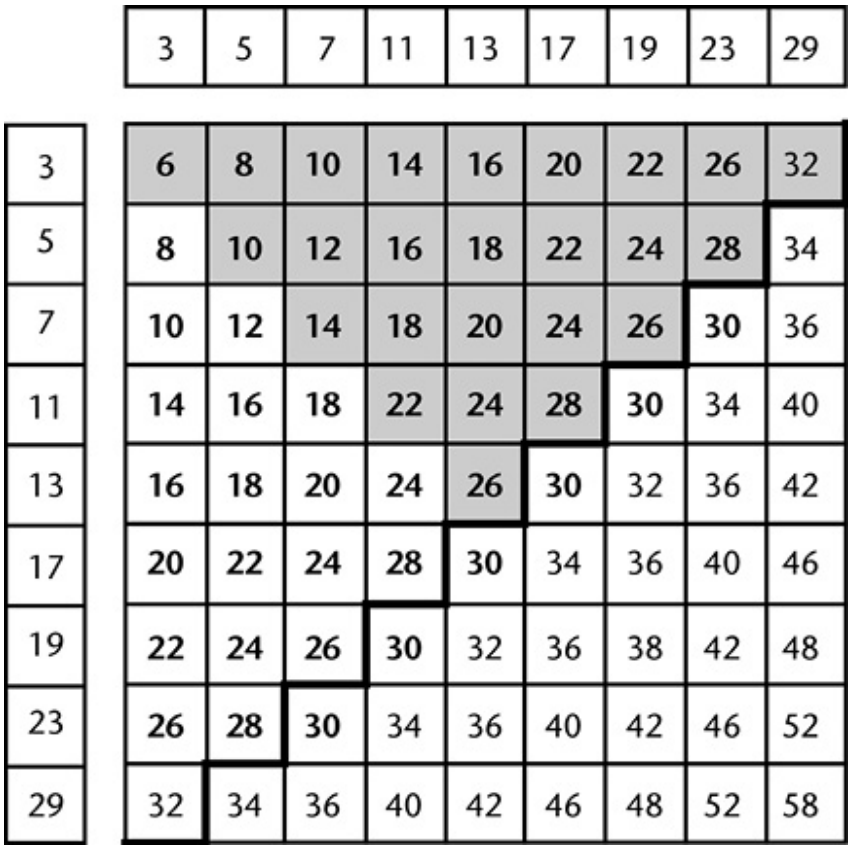
Główną przeszkodą na drodze do udowodnienia hipotezy Goldbacha jest to, że stanowi ona połączenie dwóch bardzo różnych własności. Liczby pierwsze definiuje się w oparciu o mnożenie, natomiast hipoteza mówi o dodawaniu. Dlatego właśnie powiązanie dowodzonego wniosku z jakąkolwiek sensowną cechą liczb pierwszych jest niezwykle trudne. Wydaje się, że nie ma tu nigdzie żadnego punktu zaczepienia. Dla dyrekcji wydawnictwa Faber & Faber wniosek taki musiał brzmieć bardzo przekonująco, skoro w 2000 roku ufundowała nagrodę w wysokości miliona dolarów za przedstawienie dowodu hipotezy Goldbacha – była to forma promocji powieści *Zabójcza hipoteza* Apóstolosa Doxiádisa. Wyznaczono bardzo krótki termin – rozwiązanie należało przedstawić do kwietnia 2002 roku. Nikt się nie zgłosił po odbiór nagrody, co raczej nie jest zaskoczeniem, wziąwszy pod uwagę fakt, że hipoteza ta czeka na dowód już od ponad 250 lat.

Hipotezę Goldbacha formułuje się często w postaci pytania związanego z dodawaniem zbiorów liczb naturalnych. Najłatwiej daje się w ten sposób ująć jego hipotezę dla liczb parzystych, ponieważ wymaga to dodawania tylko *dwóch* zbiorów. Aby przeprowadzić takie dodawanie, należy dodać każdą liczbę z pierwszego zbioru do każdej liczby ze zbioru drugiego, a wynikiem jest zbiór takich sum. Na przykład suma zbiorów {1, 2, 3} i {4, 5} to: 1 + 4, 2 + 4, 3 + 4, 1 + 5, 2 + 5, 3 + 5, czyli zbiór: {5, 6, 7, 8}. Niektóre elementy zbioru mogą pojawić się kilkakrotnie, na przykład 6 = 2 + 4 = 1 + 5. Tego typu powtórzenia będziemy nazywali „nadmiarowymi”.

Możemy teraz sformułować na nowo hipotezę Goldbacha dla liczb parzystych: jeśli dodamy zbiór liczb pierwszych do samego siebie, to

w wyniku otrzymamy zbiór wszystkich liczb parzystych większych od 2. Taka zmiana sformułowania może się wydawać trywialna – i taka jest – ale pozwala przenieść nasz problem do dziedziny, w której obowiązuje kilka silnych ogólnych twierdzeń. Liczba 2 trochę nam tutaj przeszkadza, ale bez trudu możemy się jej pozbyć. Jest to jedyna parzysta liczba pierwsza, jeśli więc dodamy ją do jakiejkolwiek innej liczby pierwszej, otrzymamy liczbę nieparzystą. Zatem z punktu widzenia hipotezy Goldbacha dla liczb parzystych możemy o 2 zapomnieć. Do przedstawienia liczby 4 potrzebne jest nam jednak wyrażenie 2 + 2, w takim razie musimy więc również ograniczyć nasze rozważania do liczb parzystych większych lub równych 6.

Przeprowadźmy proste doświadczenie i rozważmy liczby parzyste mniejsze lub równe 30. W tym przedziale występuje dziewięć liczb pierwszych: {3, 5, 7, 11, 13, 17, 19, 23, 29}. Wynik ich dodawania do siebie pokazano na rycinie 3 – na diagramie zaznaczono pogrubioną czcionką sumy o wartości mniejszej lub równej 30 (przedział liczb parzystych zawierający wszystkie liczby pierwsze do 29). Od razu dają się zauważyć dwie proste prawidłowości. Cała tablica jest symetryczna względem przekątnej, ponieważ $a + b = b + a$. Liczby pogrubione plasują się mniej więcej w górnym lewym rogu tabeli, nad pogrubioną linią (przekątną). Można też zwrócić uwagę na to, że w środku wychodzą one poza zaznaczoną linię. Przyczyną tego jest fakt, że dużych liczb pierwszych jest mniej niż mniejszych. Dodatkowy obszar wychodzący poza linię rekompensują z nawiązką dwie wartości 32 w prawym górnym i lewym dolnym rogu.



	3	5	7	11	13	17	19	23	29
3	**6**	**8**	**10**	**14**	**16**	**20**	**22**	**26**	32
5	**8**	**10**	**12**	**16**	**18**	**22**	**24**	**28**	34
7	**10**	**12**	**14**	**18**	**20**	**24**	**26**	**30**	36
11	**14**	**16**	**18**	**22**	**24**	**28**	**30**	34	40
13	**16**	**18**	**20**	**24**	**26**	**30**	32	36	42
17	**20**	**22**	**24**	**28**	**30**	34	36	40	46
19	**22**	**24**	**26**	**30**	32	36	38	42	48
23	**26**	**28**	**30**	34	36	40	42	46	52
29	32	34	36	40	42	46	48	52	58

Ryc. 3. Sumy par liczb pierwszych z przedziału od 3 do 30. Pogrubioną czcionką zaznaczono sumy mniejsze lub równe 30. Linia pogrubiona oznacza przekątną. Obszar zaznaczony szarym kolorem powstał po odrzuceniu par symetrycznych. Zajmuje on nieco ponad jedną czwartą powierzchni kwadratu.

Podajmy teraz kilka ogólnych oszacowań. Można to zrobić dokładniej, ale dla naszych celów wystarczą wartości przybliżone. Tablica zawiera $9 \times 9 = 81$ pól. Mniej więcej połowa liczb zapisanych w tych kratkach znajduje się w trójkącie położonym na lewo od zaznaczonej przekątnej. Z racji symetrii dodawania liczby te występują w parach, z wyjątkiem liczb położonych wzdłuż przekątnej, a więc całkowita liczba niezależnych pól wynosi około 81/4, czyli w przybliżeniu 20. W przedziale od 6 do 30 znajduje się 13 liczb parzystych. Zatem owe 20 (i ciut więcej) sum zaznaczonych pogrubioną czcionką musi trafić zaledwie w 13 liczb parzystych.

Mamy więcej potencjalnych sum liczb pierwszych z danego przedziału, niż występuje w nim liczb parzystych. To tak, jakbyśmy mogli na jarmarku rzucić 20 piłeczkami w 13 kokosów. Mamy duże szanse na strącenie wszystkich orzechów. Nie należy jednak wykluczyć, że mimo to moglibyśmy spudłować. Niektórych liczb parzystych może nie być wśród naszych sum.

W tym przykładzie tak nie jest, ale tego rodzaju argumentacja nie pozwala na wykluczenie takiej możliwości. Możemy jednak z niej wywnioskować, że musi występować duży nadmiar – wśród liczb zaznaczonych pogrubioną czcionką w analizowanej ćwiartce tabeli niektóre wartości muszą występować kilka razy. Dlaczego? Ponieważ 20 sum musi się zmieścić w zbiorze liczącym tylko 13 elementów. Zatem średnio rzecz biorąc, każda pogrubiona liczba pojawia się około 1,5 razy. (Tak naprawdę sum jest 27, a zatem lepszym oszacowaniem będzie stwierdzenie, że każda pogrubiona liczba występuje dwukrotnie). Jeśli wśród sum brakuje jakichś liczb parzystych, to ten nadmiar musi być jeszcze większy.

Możemy powtórzyć to samo ćwiczenie dla liczb z większego zakresu – powiedzmy do miliona. Ze wzoru uzyskanego na podstawie twierdzenia o liczbach pierwszych (powiemy o nim w rozdziale 9) wynika proste oszacowanie zawartości liczb pierwszych w dowolnym przedziale od 1 do *x*. Takich liczb jest $x/\ln x$. W naszym przykładzie wartość tego oszacowania wynosi w przybliżeniu 72 380. (W rzeczywistości liczb pierwszych w tym przedziale jest dokładnie 78 497). Jak zauważyliśmy, obszar zaznaczony szarym kolorem zajmuje około jednej czwartej powierzchni tabeli, a więc teraz znajduje się w nim około $n^2/4$= 250 miliardów pogrubionych liczb – sum dwóch liczb pierwszych z tego zakresu. To znacznie więcej od liczby liczb parzystych w tym przedziale, których jest pół miliona. Teraz nadmiar musi być gigantyczny – każda suma występuje w tabeli średnio 500 000 razy. Zatem szanse na to, że nie ma wśród tych sum jakiejś określonej liczby parzystej, są zdecydowanie mniejsze.

Wkładając w te rozważania nieco wysiłku, możemy je przekształcić w oszacowanie prawdopodobieństwa tego, że jakaś liczba parzysta

z danego przedziału nie jest sumą dwóch liczb pierwszych, przy założeniu, że liczby pierwsze mają losowy rozkład i występują z częstością zgodną z twierdzeniem o liczbach pierwszych – czyli że w przedziale od 1 do *x* jest ich $x/\ln x$. Tak zrobili Hardy i Littlewood. Zdawali sobie sprawę, że takie podejście nie jest ścisłe, ponieważ liczby pierwsze mają precyzyjną definicję i nie są wielkością losową. Niemniej należy oczekiwać, że prawdziwe wyniki powinny być zgodne z wnioskami płynącymi z takiego modelu probabilistycznego, ponieważ definicja liczb pierwszych ma niewielki związek z tym, co się dzieje, gdy dodajemy je do siebie.

W tym zakresie uczeni opracowali jeszcze kilka innych podobnych podejść, starając się jednak zachować ścisłą argumentację. Przykładem tego mogą być różne odmiany metody sita będące rozszerzeniem sita Eratostenesa. Innym użytecznym narzędziem są ogólne twierdzenia mówiące o gęstości liczb w sumach dwóch zbiorów – czyli o proporcjach pojawiających się liczb w miarę wzrostu rozmiaru zbiorów.

Gdy jakaś hipoteza matematyczna okazuje się ostatecznie poprawna, proces ten przebiega zwykle według takiego samego, standardowego scenariusza. Przez pewien czas uczeni dowodzą poprawności hipotezy przy pewnych określonych ograniczeniach. Każdy kolejny taki dowód poprawia poprzedni wynik, usuwając jakieś ograniczenie, ale ostatecznie cały proces traci impet. W końcu jednak pojawia się jakaś nowa, genialna idea, która pozwala dokończyć dowód.

Załóżmy na przykład, że istnieje hipoteza teorii liczb głosząca, że każdą dodatnią liczbę całkowitą można przedstawić w jakiś sposób z wykorzystaniem, powiedzmy, sześciu specjalnych wartości (liczb pierwszych, kwadratów, sześcianów czy dowolnych innych konstrukcji). W tym wypadku znaczenie ma to, że mówimy o *każdej* dodatniej liczbie całkowitej i *sześciu* specjalnych liczbach. Pierwsze próby udowodnienia takiej hipotezy prowadzą do uzyskania dowodów dla dużo słabszych założeń, ale w kolejnych krokach wynik powoli się poprawia.

Pierwszym krokiem jest często przedstawienie dowodu twierdzenia w rodzaju: każdą dodatnią liczbę całkowitą niepodzielną przez 3 i 11, z wyjątkiem jakiejś skończonej ich liczby, można przedstawić za pomocą określonej olbrzymiej kombinacji liczb specjalnych – powiedzmy, że będzie ich 10^{666}. Z takiego twierdzenia zazwyczaj nie wynika, ile jest takich wyjątków, uzyskanego wyniku nie można więc zastosować bezpośrednio do określonej liczby całkowitej. Następny krok polega na wyznaczeniu konkretnej granicy – na przykład na udowodnieniu, że można w ten sposób przedstawić każdą liczbę całkowitą większą od $10^{10^{42}}$. Później uczeni pozbywają się wymagania na podzielność przez 3, a w kolejnym kroku udaje się usunąć wymaganie na podzielność przez 11. Potem matematycy zmniejszają wartość 10^{666} lub $10^{10^{42}}$, a może nawet obie. Typowym krokiem w tym kierunku może być na przykład prezentacja dowodu twierdzenia, że każdą liczbę całkowitą większą od $5{,}8 \times 10^{17}$ można przedstawić za pomocą 4298 liczb specjalnych.

Tymczasem inni badacze posuwają się w przeciwnym kierunku, od małych liczb do dużych, często wykorzystując do tego komputery, i dowodzą, powiedzmy, że każdą liczbę mniejszą lub równą 10^{12} można przedstawić za pomocą najwyżej sześciu liczb specjalnych. W ciągu następnego roku różnym naukowcom lub ich grupom udaje się pięciokrotnie przesunąć wartość 10^{12} i ostatecznie w twierdzeniu widnieje liczba $11{,}0337 \times 10^{29}$. Wprowadzenie takich poprawek nie było ani rutynowe, ani łatwe. Za każdym razem wymagało to użycia specjalnych, skomplikowanych technik, z których nie można wyprowadzić żadnego ogólniejszego podejścia i każdy kolejny dowód jest coraz dłuższy i bardziej złożony. Przez kilka lat badacze wprowadzają kolejne poprawki tego rodzaju, stosują te same ogólne idee, ale z wykorzystaniem silniejszych komputerów i nowych usprawnień. W efekcie wartość w twierdzeniu zwiększa się do 10^{43}. Jednak w końcu możliwości tego podejścia wyczerpują się i wszyscy się zgadzają, że żadne dalsze usprawnienia metody nie doprowadzą do udowodnienia pełnej hipotezy.

W tym momencie hipoteza odchodzi w niebyt, ponieważ nikt już nad nią nie pracuje. W niektórych wypadkach prace ulegają całkowitemu zatrzymaniu. Czasami zaś następuje dwadzieścia lat kompletnej ciszy… po czym niespodzianie, zupełnie znikąd, Cheesberger i Fryteck oznajmiają, że po sformułowaniu hipotezy z wykorzystaniem zespolonych metaergodycznych kwazistogów i zastosowaniu bizantyjskiej teorii kolaboranckiej udało im się uzyskać pełny dowód. Po kilku latach sporów na temat konkretnych przejść w wywodzie logicznym i wypełnieniu kilku luk społeczność matematyków uznaje, że przedstawiony dowód jest poprawny, i od razu wszyscy zaczynają się zastanawiać, czy istnieje jakiś lepszy sposób uzyskania tego samego wyniku lub może nawet udowodnienia ogólniejszej wersji hipotezy.

W kolejnych rozdziałach nieraz przekonamy się, że postęp prac nad omawianymi zagadnieniami przebiegał właśnie według takiego scenariusza. Ponieważ podobne opowieści stają się nużące – bez względu na to, jak dumni są Buggins i Krumm z tego, że udało im się poprawić wykładnik w hipotezie Jekylla i Hyde'a z 1,773 na 1,771 + ε, dla dowolnego ε większego od zera – przedstawimy w takich przypadkach tylko kilka charakterystycznych wyników, pomijając całą resztę. Nie oznacza to, że osiągnięcie Bugginsa i Krumma było nieważne. Mogło ono mieć kluczowe znaczenie na drodze do wielkiego dowodu Cheesbergera i Frytecka. Jednak tylko specjaliści śledzący na bieżąco rozwój wypadków czekają z zapartym tchem na kolejny, najmniejszy nawet krok naprzód.

W dalszych rozdziałach nie będziemy omawiali przebiegu wydarzeń aż tak szczegółowo, zobaczmy jednak teraz, jak wyglądał ten proces dla hipotezy Goldbacha.

Zacznijmy od twierdzeń, które przybliżają nas do przedstawienia pełnego dowodu hipotezy Goldbacha. Pierwszy przełom nastąpił w 1923 roku, gdy Hardy i Littlewood za pomocą opracowanych przez siebie technik analitycznych udowodnili hipotezę Goldbacha dla liczb nieparzystych dla wszystkich odpowiednio dużych liczb. Ich dowód bazuje jednak na innej wielkiej hipotezie – na uogólnionej hipotezie

Riemanna – którą omówimy w rozdziale 9. Jest to problem wciąż otwarty, zatem w rozwiązaniu Hardy'ego i Littlewooda występowała istotna luka. W 1930 roku Lew Sznirelman usunął tę lukę, wykorzystawszy wymyślną wersję ich wywodu, opartą na metodach odsiewania liczb. Udowodnił, że niezerowy odsetek wszystkich liczb można przedstawić jako sumę dwóch liczb pierwszych. Łącząc ten wynik z ogólnymi twierdzeniami na temat dodawania ciągów, dowiódł, że musi istnieć pewna liczba C taka, że każda liczba naturalna większa od 1 jest sumą najwyżej C liczb pierwszych. Liczbę tę nazywa się stałą Sznirelmana. Iwan Matwiejewicz Winogradow uzyskał podobny wynik w 1937 roku, ale jego metoda również nie precyzowała, co znaczy „odpowiednio duża liczba". W 1939 roku K. Borozdin dowiódł, że nie może to być wartość większa niż $3^{14\,348\,907}$. Do roku 2002 Liu Ming-Chit i Wang Tian-Ze zmniejszyli tę „górną granicę" do wartości e^{3100}, czyli około 2×10^{1346}. To o wiele mniej, ale wciąż zbyt dużo, by pozostałe, mniejsze liczby można było sprawdzić na komputerze.

W 1969 roku N.I. Klimow wyznaczył pierwsze ograniczenie na wartość stałej Sznirelmana – ustalił, że nie może być ona większa niż 6 miliardów. Innym matematykom udało się później obniżyć znacznie tę wartość i w 1982 roku Hans Riesel i Robert Vaughan zmniejszyli ją do 19. Chociaż 19 to zdecydowanie lepiej niż 6 miliardów, dowody wskazywały, że stała Sznirelmana powinna wynosić zaledwie 3. W 1995 roku Leszek Kaniecki zmniejszył górną granicę do wartości 6, przy założeniu, że każda liczba pierwsza jest sumą najwyżej pięciu liczb pierwszych, ale w tym celu musiał również przyjąć prawdziwość hipotezy Riemanna. Uzyskany przez niego wynik, w połączeniu z rachunkowym sprawdzeniem hipotezy Riemanna do wartości 4×10^{14}, którego dokonał Jörg Richstein, pozwolił udowodnić, że stała Sznirelmana wynosi najwyżej 4, oczywiście przy założeniu poprawności hipotezy Riemanna. W 1997 roku Jean-Marc Deshouillers, Gove Effinger, Herman te Riele i Dmitrij Zinowiew pokazali, że z prawdziwości uogólnionej hipotezy Riemanna wynika prawdziwość hipotezy Goldbacha dla liczb nieparzystych. To znaczy,

że każda liczba nieparzysta z wyjątkiem 1, 3 i 5 jest sumą trzech liczb pierwszych.

Ponieważ hipoteza Riemanna na razie pozostaje nieudowodniona, warto byłoby pozbyć się tego założenia. W 1995 roku francuski matematyk Olivier Ramaré pokazał bez odwoływania się do hipotezy Riemanna, że liczby nieparzyste można przedstawić jako sumę najwyżej siedmiu liczb pierwszych. Tak naprawdę udowodnił coś silniejszego – że każda liczba parzysta jest sumą najwyżej sześciu liczb pierwszych. (Aby zastosować to twierdzenie do liczby nieparzystej, należy od niej odjąć 3: wynikiem będzie liczba parzysta, a więc możemy ją zapisać jako sumę najwyżej sześciu liczb pierwszych. Nasza wyjściowa liczba jest równa tej sumie plus 3 – kolejna liczba pierwsza – a zatem do jej zapisu potrzeba najwyżej siedmiu liczb pierwszych). Główny przełom nastąpił, gdy udało się poprawić oszacowania na ilość liczb z określonego przedziału będących sumą dwóch liczb pierwszych. Najważniejszy wynik na tym polu badań uzyskał Ramaré – pokazał, że dla każdej liczby n większej od e^{67} (czyli około $1{,}25 \times 10^{29}$) przynajmniej jedna piąta liczb z przedziału między n i $2n$ jest sumą dwóch liczb pierwszych. Jeśli wykorzysta się metody przesiewowe w powiązaniu z twierdzeniem Hansa Heinricha Ostmanna o sumach ciągów, poprawionym przez Deshouillersa, to można udowodnić, że każda liczba większa od 10^{30} jest sumą najwyżej sześciu liczb pierwszych.

Uczeni musieli się teraz rozprawić jeszcze z luką między wartością 4×10^{14}, do której Jörg Richstein sprawdził poprawność twierdzenia z wykorzystaniem komputera, a 10^{30}. Jak to często bywa w takich wypadkach, liczby te są zbyt duże, by dało się je sprawdzić bezpośrednio na komputerze, Ramaré więc udowodnił serię specjalnych twierdzeń na temat ilości liczb pierwszych w niewielkich przedziałach. Poprawność tych dowodów zależy od prawdziwości hipotezy Riemanna dla liczb z przedziałów nieprzekraczających pewnych konkretnych wartości, co można sprawdzić za pomocą komputera. Zatem dowód składa się głównie z wywodów pojęciowych, popartych obliczeniami komputerowymi dla tych

konkretnych sytuacji. Ramaré kończy swój artykuł uwagą, że w zasadzie podobne podejście powinno pozwolić na obniżenie wymagania na ilość liczb pierwszych z 7 do 5. Próbując jednak tego dokonać, natrafił na poważne przeszkody i, jak napisał, takiego dowodu „nie można przeprowadzić z wykorzystaniem dostępnych obecnie komputerów".

W 2012 roku Terence Tao pokonał te przeszkody dzięki zastosowaniu nowych, zupełnie innych pomysłów. Tao zamieścił już w Internecie artykuł opisujący uzyskany wynik – w chwili gdy piszę te słowa, artykuł jest jeszcze w recenzji[12]. Główne twierdzenie artykułu brzmi: każda liczba nieparzysta jest sumą najwyżej pięciu liczb pierwszych. Zmniejsza to stałą Sznirelmana do wartości 6. Tao jest znany z tego, że udaje mu się rozwiązywać trudne problemy z różnych działów matematyki. W swoim dowodzie wykorzystał kilka zaawansowanych technik i musiał się posłużyć komputerem. Gdyby liczbę 5 z twierdzenia Tao udało się zmniejszyć do wartości 3, oznaczałoby to udowodnienie hipotezy Goldbacha dla liczb nieparzystych i zmniejszenie ograniczenia na wartość stałej Sznirelmana do wartości 4. Tao sądzi, że powinno być to możliwe, choć uzyskanie takiego wyniku będzie wymagało zastosowania jakichś nowych idei.

Hipoteza Goldbacha dla liczb parzystych wydaje się jeszcze trudniejsza. W 1998 roku Jean-Marc Deshouillers, Yannick Saouter i Herman te Riele sprawdzili ją dla wszystkich liczb parzystych do wartości 10^{14}. W 2007 roku Tomás Oliveira e Silva doszedł już do wartości 10^{18} i wciąż liczy dalej. Wiemy, że każda parzysta liczba całkowita jest sumą najwyżej sześciu liczb pierwszych – udowodnił to Ramaré w 1995 roku. W 1973 roku Chen Jing-Run dowiódł, że każda odpowiednio duża parzysta liczba całkowita jest sumą liczby pierwszej i półpierwszej (czyli liczby, która jest albo liczbą pierwszą, albo iloczynem dwóch liczb pierwszych). To już blisko, ale na otwarcie szampana jeszcze za wcześnie. Tao oświadczył, że jego metody nie nadają się do prac nad hipotezą Goldbacha dla liczb parzystych. Dodanie do siebie trzech liczb pierwszych prowadzi do powstania dużo większego nadmiaru – w rozumieniu takim, jakie

rozważaliśmy przy okazji omawiania ryciny 3 – niż zsumowanie dwóch liczb pierwszych w przypadku hipotezy Goldbacha dla liczb parzystych, i metody Tao i Ramarégo wielokrotnie ten fakt wykorzystują.

Za kilka lat możemy się więc doczekać pełnego dowodu hipotezy Goldbacha dla liczb nieparzystych, z którego w szczególności będzie wynikało, że każda liczba parzysta jest sumą najwyżej czterech liczb pierwszych. Jednak nawet wtedy hipoteza Goldbacha dla liczb parzystych będzie nas wciąż tak samo zdumiewała jak kiedyś Eulera i Goldbacha.

W ciągu 2300 lat, jakie upłynęły od czasu, gdy Euklides udowodnił kilka podstawowych twierdzeń o liczbach pierwszych, udało nam się wiele dowiedzieć na temat tych niezwykle ważnych, choć nieuchwytnych liczb. Jednak im więcej się dowiadujemy, tym wyraźniej widzimy długą listę tego, co pozostaje wciąż do wyjaśnienia.

Wiemy na przykład, że istnieje nieskończenie wiele liczb pierwszych postaci $4k + 1$ i $4k + 3$, czy też mówiąc ogólniej, że każdy ciąg arytmetyczny[13] postaci $ak + b$, dla stałych wartości a i b, zawiera nieskończenie wiele liczb pierwszych, pod warunkiem że a i b nie mają wspólnego dzielnika. Załóżmy na przykład, że $a = 18$. Wtedy $b = 1, 5, 7, 11, 13, 17$. Zatem istnieje nieskończenie wiele liczb pierwszych postaci $18k + 1$, $18k + 5$, $18k + 7$, $18k + 11$, $18k + 13$ oraz $18k + 17$. Nie jest to prawdą dla, powiedzmy, $18k + 6$, ponieważ 18 jest wielokrotnością 6. Żaden ciąg arytmetyczny nie może się składać *wyłącznie* z liczb pierwszych, ale z jednego z najnowszych ważnych dokonań, z twierdzenia Greena i Tao, wynika, że zbiór liczb pierwszych zawiera dowolnie długie ciągi arytmetyczne. Dowód tego, przedstawiony w 2004 roku przez Bena Greena i Terence'a Tao, jest skomplikowany i głęboki. Tchnie jednak nadzieją: pokazuje, że trudne otwarte problemy, nawet te, które wydają się nieprzeniknione, udaje się czasami rozwiązać.

Wczuwając się w rolę specjalistów od algebry, możemy od razu spróbować wymyślić więcej skomplikowanych wzorów z k. Nie ma

liczb pierwszych postaci k^2i $k^2 - 1$ (w tym drugim wypadku z wyjątkiem liczby 3), ponieważ wyrażenia te rozkładają się na czynniki pierwsze. Jednak wyrażenie $k^2 + 1$ nie ma żadnego oczywistego rozkładu na czynniki i możemy podać wiele liczb pierwszych tego rodzaju:

$$2 = 1^2 + 1$$
$$5 = 2^2 + 1$$
$$17 = 4^2 + 1$$
$$37 = 6^2 + 1$$

i tak dalej. Przykładem większej liczby tego rodzaju może być:

$$18\ 672\ 907\ 718\ 657 = (4\ 321\ 216)^2 + 1.$$

Istnieje hipoteza, że takich liczb pierwszych jest nieskończenie wiele, ale jak dotąd nie udało się udowodnić żadnego takiego stwierdzenia dla jakiegokolwiek konkretnego wielomianu, w którym *k* byłoby podniesione do potęgi większej niż jeden. Wiktor Buniakowski wysunął w 1857 roku bardzo prawdopodobnie brzmiącą hipotezę: każdy wielomian, w którym *k* nie ma oczywistych dzielników, opisuje nieskończenie wiele liczb pierwszych. Do wspomnianych tu wyjątków należą nie tylko wielomiany rozkładalne, ale również takie jak $k^2 + k + 2$, który zawsze można podzielić przez 2, mimo że nie ma czynników algebraicznych.

Wydaje się, że niektóre wielomiany mają szczególne własności. Klasycznym przykładem jest $k^2 + k + 41$, który jest liczbą pierwszą dla *k* = 0, 1, 2, …, 40, a także dla *k* = –1, –2, …, –40. Długie ciągi liczb pierwszych dla kolejnych wartości *k* należą do rzadkości i udało nam się już zebrać na ich temat pewną wiedzę. Jednak cały ten obszar pozostaje wciąż tajemniczy.

Niemal równie słynna jak hipoteza Goldbacha, i najwyraźniej równie trudna, jest hipoteza o liczbach bliźniaczych. Głosi ona, iż

istnieje nieskończenie wiele par liczb pierwszych takich, że ich różnica wynosi 2. Przykładami mogą być:

3, 5 5, 7 11, 13 17,19.

Największe znane liczby bliźniacze (stan na styczeń 2012 roku) to:

$$3\,756\,801\,695\,685 \times 2^{666\,669} \pm 1.$$

W pełnym zapisie każda z tych liczb ma 200 700 cyfr. Odkryto je w 2011 roku w ramach rozproszonego projektu obliczeniowego PrimeGrid. W 1915 roku Viggo Brun wykorzystał pewną odmianę sita Eratostenesa do udowodnienia, że w przeciwieństwie do szeregu odwrotności wszystkich liczb pierwszych szereg odwrotności wszystkich liczb bliźniaczych jest zbieżny. Zatem w tym sensie liczby bliźniacze są względnie rzadkie. Wykorzystując podobną metodę, Brun dowiódł również, że istnieje nieskończenie wiele liczb całkowitych *n* takich, że *n* i *n* + 2 mają najwyżej dziewięć czynników pierwszych. Hardy i Littlewood posłużyli się metodami heurystycznymi do udowodnienia, że liczba liczb bliźniaczych mniejszych lub równych wartości *x* powinna zbliżać się asymptotycznie do

$$2a\frac{n}{(\ln n)^2}$$

gdzie *a* jest stałą o wartości wynoszącej około 0,660161. Ich dowód opiera się na założeniu, że liczby pierwsze występują w sposób losowy z częstością taką, iż liczba liczb pierwszych o wartości nieprzekraczającej *x* jest w przybliżeniu równa $x/\ln x$. Istnieje wiele podobnych hipotez i wzorów heurystycznych, ale tak jak poprzednio, brak ścisłych dowodów.

Z liczbami pierwszymi wiążą się setki otwartych pytań. Część z nich to tylko ciekawostki, ale inne opisują głębokie i ważne problemy. Z niektórymi z nich spotkamy się jeszcze w rozdziale 9. Mimo ogromnych postępów, jakie poczynili matematycy w ciągu minionych

dwóch i pół tysiąca lat, skromne liczby pierwsze nie straciły nic ze swego uroku i tajemniczości.

9 Algorytm Agrawala, Kayali i Saxeny (test AKS) ma następującą postać:

Dana wejściowa: liczba naturalna *n*.
1. Jeśli *n* jest potęgą dowolnej mniejszej liczby, wypisz wynik: „Liczba złożona" i zakończ działanie algorytmu.
2. Znajdź najmniejszą liczbę *r* taką, że najmniejsza potęga *r*, która jest równa 1 w arytmetyce z modułem *n*, wynosi przynajmniej $(\ln n)^2$.
3. Jeśli dowolna liczba mniejsza lub równa *r* ma wspólny dzielnik z *n*, wypisz wynik: „Liczba złożona" i zakończ działanie algorytmu.
4. Jeśli *n* jest mniejsze lub równe *r*, wypisz wynik: „Liczba pierwsza" i zakończ działanie algorytmu.
5. Dla wszystkich liczb naturalnych *a* zmieniających się od 1 do określonej wartości sprawdź, czy wielomian $(x+a)^n$ jest taki sam jak $x^n + a$ w arytmetyce z modułem *n* i w arytmetyce z modułem $x^r - 1$. Jeśli taka równość jest spełniona dla dowolnej z liczb, wypisz wynik: „Liczba złożona" i zakończ działanie algorytmu.
6. Wypisz wynik: „Liczba pierwsza".

10 Mam tu na przykład na myśli wzór $\lfloor A^{3^n} \rfloor$, w którym symbol przypominający nawiasy kwadratowe oznacza liczbę całkowitą mniejszą lub równą liczbie zapisanej w środku. W 1947 roku William H. Mills udowodnił, iż istnieje pewna stała rzeczywista *A* taka, że dla dowolnego *n* wynik tego wzoru jest liczbą pierwszą. Zakładając prawdziwość hipotezy Riemanna, można wyliczyć, że najmniejszą wartością *A*, która ma taką własność, jest liczba wynosząca w przybliżeniu 1,306. Problem jednak w tym, że stałą tę definiuje się za pomocą odpowiedniego ciągu liczb pierwszych i podany przed chwilą wzór jest jedynie symbolicznym zapisem takiego ciągu. Więcej podobnych wzorów, w tym nawet takich, które obejmują wszystkie liczby pierwsze, można znaleźć na stronach internetowych w języku angielskim: http://mathworld.wolfram.com/PrimeFormulas.html oraz http://en.wikipedia.org/wiki/Formula_for_primes.

11 Jeśli *n* jest nieparzyste, to $n - 3$ jest parzyste, a jeżeli *n* jest większe od 5, to $n - 3$ jest większe od 2. Z pierwszej hipotezy wynika, że $n - 3 = p + q$, a zatem $n = p + q + 3$.

12 Artykuł przeszedł pomyślnie proces recenzji i został ostatecznie opublikowany 24 czerwca 2013 roku: Terence Tao, *Every odd number greater than 1 is the sum of at most five primes* (*Każda liczba nieparzysta większa od 1 jest sumą najwyżej pięciu liczb pierwszych*), „Mathematics of Computation" 2013, DOI: http://dx.doi.org/10.1090/S0025–5718–2013–02733–0 (przyp. tłum.).

13 Osobiście wolę to określenie od bardziej staroświeckiego, ale może lepiej znanego terminu „postęp arytmetyczny". Nikt już nie mówi o „postępach", z wyjątkiem postępu arytmetycznego i geometrycznego. Pora pójść z duchem czasu.

3. Zagadkowa liczba π

Kwadratura koła

Liczby pierwsze są starym pojęciem, ale koło jest jeszcze starsze. Z kołem związany jest wielki problem matematyczny, którego rozwiązanie zajęło uczonym dwa tysiące lat. Problem ten należy do całej grupy blisko ze sobą związanych zagadnień geometrycznych, które odziedziczyliśmy po starożytnych filozofach. Głównym bohaterem tej opowieści jest liczba π, którą wszyscy poznajemy w szkole przy okazji omawiania kół i kul. Jej wartość wynosi nieco ponad 3,14159. Często też używa się przybliżenia 22/7. Cyfry liczby π ciągną się po przecinku w nieskończoność, nie tworząc przy tym żadnego powtarzającego się ciągu. Bieżący rekord na obliczenie wartości liczby π, wynoszący 10 bilionów cyfr, należy do Alexandra Yee i Shigeru Kondo. Rekord ten padł w październiku 2011 roku[14]. Tego typu obliczenia są dobrym sposobem na sprawdzenie szybkości komputerów lub opracowanie i przetestowanie nowych metod wyznaczania liczby π, ale nie mają one większego matematycznego znaczenia. Liczba π jest interesująca nie ze względu na to, że umożliwia obliczenie obwodu koła, ale dlatego, że pojawia się praktycznie we wszystkich gałęziach matematyki – nie tylko we wzorach związanych z kołami i kulami – i prowadzi nas na bardzo głęboką wodę. Szkolne wzory są jednak ważne, ponieważ pozwalają nam poznać rodowód tej tajemniczej liczby, której początki sięgają geometrii starożytnych Greków.

W tamtym okresie jednym z największych wyzwań był nierozwiązany problem kwadratury koła. O „kwadraturze koła" mówimy często potocznie, gdy mamy na myśli coś niemożliwego do

osiągnięcia – jakbyśmy wyobrażali sobie, że ktoś próbuje wetknąć kwadratowy kołek w okrągły otwór. Jak wiele frazeologizmów wywodzących się z nauk ścisłych, znaczenie tego powiedzenia zmieniło się z upływem czasu. W starożytnej Grecji kwadratura koła była całkowicie sensownym pojęciem. Różnica w kształcie dwóch figur – to, czy są kanciaste, czy zakrzywione – nie ma tu najmniejszego znaczenia: wiele podobnych problemów doczekało się poprawnego rozwiązania[15]. Ostatecznie okazało się jednak, że tego konkretnego problemu nie można rozwiązać przy zachowaniu przyjętych ograniczeń. Dowiedzenie tego wymagało dużego sprytu i znajomości zaawansowanych technik matematycznych, ale ogólny sens całego wywodu nie jest trudny do zrozumienia.

Mówiąc o kwadraturze koła, matematycy mają na myśli problem skonstruowania kwadratu, którego *pole* będzie równe polu danego koła, przy czym w konstrukcji można korzystać wyłącznie ze standardowych metod euklidesowych. W starożytnej Grecji dopuszczano tak naprawdę stosowanie w geometrii również innych metod, zatem jeden z aspektów problemu polega na dokładnym ustaleniu, które są dopuszczalne. W tym kontekście niemożność rozwiązania tak postawionego zadania wynika zatem z ograniczeń owych metod – nie oznacza to, że nie potrafimy obliczyć pola koła. Musimy po prostu znaleźć inne podejście. Dowód nierozwiązywalności tego problemu wyjaśnia także, dlaczego greckim matematykom i ich następcom nie udało się znaleźć konstrukcji wymaganego rodzaju: taka konstrukcja po prostu nie istnieje. Z perspektywy czasu rozumiemy też, dlaczego dawni matematycy musieli się uciekać do bardziej zawiłych metod. Widzimy więc, że rozwiązanie tego problemu, mimo że sprowadza się do udowodnienia niewykonalności postawionego zadania, pozwoliło nam wyjaśnić wielką historyczną zagadkę. Dzięki temu ludzie przestali marnotrawić czas na poszukiwanie nieistniejącej konstrukcji – z wyjątkiem nielicznych zatwardziałych zapaleńców, do których nie dociera, że jest to niemożliwe, bez względu na to, jak prosto próbuje im się to wyjaśnić[16].

W *Elementach* Euklides stosuje tradycyjne metody konstrukcji figur geometrycznych, które polegają na użyciu wyidealizowanych wersji dwóch przyrządów matematycznych: linijki i cyrkla. Za pomocą takich przyrządów matematycy „rysują" diagramy na pojęciowej kartce papieru, którą jest płaszczyzna euklidesowa.

Z budowy tych przyrządów wynika, co można nimi narysować. Cyrkiel składa się z dwóch sztywnych ramion połączonych ruchomym zawiasem. W jednym z końców montuje się ostry szpikulec, w drugim – zaostrzony ołówek. Używa się go do rysowania okręgów, lub ich fragmentów, o określonym środku i promieniu. Linijka jest prostszym przyrządem: ma prostą krawędź, która umożliwia rysowanie odcinków linii prostej. W przeciwieństwie do linijek, które można kupić w sklepie papierniczym, na linijkach Euklidesa nie ma żadnych oznaczeń – jest to ważne ograniczenie, decyduje bowiem o tym, co można za ich pomocą zrobić.

Nietrudno zauważyć, w jakim sensie linijka i cyrkiel używane w konstrukcjach geometrycznych są idealizacjami: przyjmuje się, że za ich pomocą można rysować nieskończenie cienkie linie. Co więcej, powstałe za ich pomocą linie są doskonale proste, a okręgi – doskonale okrągłe. Papier jest zaś doskonale płaski i równy. Kolejnym kluczowym składnikiem geometrii euklidesowej jest pojęcie punktu – to kolejna idealizacja. Punkt nie jest kropką na papierze, ale fizyczną niemożliwością: obiektem bez rozmiaru. „Punkt – powiada Euklides w pierwszym zdaniu *Elementów* – jest tym, co nie ma części". Brzmi to trochę jak definicja atomu albo, dla kogoś obeznanego ze współczesną fizyką, definicja cząstki elementarnej, ale w porównaniu z geometrycznym punktem obiekty te są gigantyczne. Jednak z naszej ludzkiej perspektywy wyidealizowany punkt Euklidesa, atom i kropka zrobiona ołówkiem na papierze są do siebie dość podobne pod względem geometrii.

Te ideały są nieosiągalne w rzeczywistym świecie. Nawet gdybyśmy niezwykle starannie wykonali przyrządy, naostrzyli ołówek i wygładzili dokładnie papier, to i tak nie osiągniemy ideału. Taki idealizm ma jednak swoje dobre strony, ponieważ pozwala znacznie uprościć matematykę. Dwie proste narysowane ołówkiem na papierze przecinają się na przykład w rozmytym obszarze o kształcie

równoległoboku, natomiast proste matematyczne przecinają się zawsze w jednym punkcie. Wnioski wysnute z rozważań na temat idealnych okręgów i prostych można często przenieść na rzeczywiste, niedoskonałe obiekty. Właśnie w ten sposób matematyka dokonuje swoich cudów.

Dwa punkty wyznaczają jedną jedyną przechodzącą przez nie prostą. Aby ją skonstruować, należy umieścić idealną linijkę tak, by przechodziła przez oba punkty, i przeciągnąć idealnym ołówkiem wzdłuż jej krawędzi. Dwa punkty wyznaczają również okrąg – wystarczy przyjąć, że jeden z nich jest jego środkiem, i wbić w to miejsce ostrze cyrkla, a następnie otworzyć go tak, by czubek ołówka znalazł się w drugim punkcie. Teraz wystarczy już tylko obrócić cyrkiel dookoła, zakreślając ołówkiem łuk, pilnując przy tym, żeby ostrze cyrkla cały czas pozostawało nieruchome. Dwie proste określają jednoznacznie jeden punkt – ten, w którym się przecinają – z wyjątkiem prostych równoległych, które się nie przecinają, ale za to otwierają szeroko puszkę Pandory z przeróżnymi problemami logicznymi. Prosta i okrąg wyznaczają jednoznacznie dwa punkty, jeśli się przecinają; jeden punkt, jeśli prosta jest styczną do okręgu; ani jednego punktu, jeśli okrąg jest zbyt mały, by mógł przeciąć prostą. Podobnie dwa okręgi mogą się przecinać w dwóch punktach, w jednym lub w żadnym.

We współczesnym ujęciu geometrii euklidesowej podstawowym pojęciem jest również odległość. Odległość między dwoma punktami mierzy się wzdłuż łączącej je linii prostej. Euklidesowi udało się opisać geometrię bez jawnego wprowadzania pojęcia odległości dzięki temu, że znalazł sposób na ustalenie, iż dwa odcinki mają taką samą długość, bez definiowania samej długości. To całkiem proste: wystarczy otworzyć cyrkiel tak, by jego końce dotykały końców jednego odcinka, a następnie przyłożyć go do drugiego odcinka i sprawdzić, czy pasuje. Jeśli tak, to oba odcinki mają taką samą długość, a jeśli nie, to są różnej długości. Takie podejście w żadnym kroku nie wymaga mierzenia ich faktycznej długości.

Z tych podstawowych składników matematycy potrafią skonstruować więcej interesujących figur i ich kombinacji. Trzy punkty nieleżące na jednej prostej definiują trójkąt. Dwie

przecinające się proste tworzą kąt. Kąt prosty ma szczególne znaczenie: prosta odpowiada połączeniu ze sobą dwóch kątów prostych. I tak dalej, i tak dalej. *Elementy* Euklidesa zawierają 13 ksiąg, w których rozważa się coraz głębsze wnioski wynikające z tych prostych założeń początkowych.

Przeważającą część *Elementów* stanowią twierdzenia – udowodnione własności geometryczne. Euklides wyjaśnia również, jak należy rozwiązywać zadania geometryczne, tworząc „konstrukcje" za pomocą linijki i cyrkla: mając dane dwa punkty i łączący je odcinek, wyznacz jego symetralną. Dokonaj trysekcji odcinka: wyznacz punkt położony dokładnie w jednej trzeciej jego odległości. Wyznacz prostą dzielącą dany kąt na dwie równe części. Jednak niektóre proste konstrukcje okazały się trudniejsze, niż można było przypuszczać. Na przykład: dokonaj trysekcji kąta, czyli innymi słowy, skonstruuj kąt o wielkości równej jednej trzeciej danego kąta. Można dokonać takiego podziału odcinka, ale nikomu nie udało się znaleźć sposobu na podzielenie w ten sposób kąta. Owszem, wymyślono metody na przybliżenie takiego podziału z dowolną dokładnością, ale nikt nie przedstawił precyzyjnej konstrukcji z wykorzystaniem linijki bez podziałki i cyrkla. Na szczęście dokładny podział kąta na trzy części nie jest tak naprawdę nikomu do niczego potrzebny, ten konkretny problem więc nie spędzał raczej matematykom snu z powiek.

Bardziej żenujący był fakt, że nikt nie potrafił przedstawić rozwiązania innego problemu: mając dane koło, skonstruuj kwadrat o takim samym polu. Chodzi oczywiście o kwadraturę koła. Z punktu widzenia geometrii starożytnych Greków brak rozwiązania tego problemu sprawiał, że nie można było twierdzić, iż koło *ma* jakieś pole, mimo że wyraźnie widać, iż zajmuje ono dobrze określony obszar, i intuicyjnie czujemy, że obszar ten musi mieć *jakieś* pole. Euklides i jego następcy, szczególnie Archimedes, wybrali rozwiązanie pragmatyczne: załóżmy, że koło ma pole, ale nie oczekujmy, że uda nam się skonstruować kwadrat o takim samym polu. W takiej sytuacji i tak można wiele powiedzieć na temat koła. Można na przykład udowodnić, przeprowadzając ścisły logiczny wywód, że pole koła jest proporcjonalne do kwadratu jego średnicy.

Bez rozwiązania kwadratury koła nie da się jednak skonstruować odcinka o długości równej stałej proporcjonalności z tej zależności.

Grekom nie udało się przeprowadzić kwadratury koła za pomocą linijki i cyrkla, wymyślili więc inne metody. Jedna polega na wykorzystaniu krzywej o nazwie kwadratrysa[17]. Późniejsi komentatorzy wyolbrzymiali znaczenie, jakie starożytni Grecy przypisywali konstrukcji z wykorzystaniem wyłącznie cyrkla i linijki, i obecnie nie jest nawet pewne, czy Grecy uważali problem kwadratury koła za aż tak istotny. W XIX stuleciu osiągnął on już jednak rangę niezwykle uciążliwej niedogodności. Matematyk niepotrafiący odpowiedzieć na tak proste pytanie był niczym szef kuchni z pięciogwiazdkowej restauracji, który nie umie ugotować jajka.

Wydaje się, że kwadratura koła jest zagadnieniem z zakresu geometrii. I słusznie, bo jest to problem geometryczny. Okazało się jednak, że do jego rozwiązania potrzebna jest znajomość nie geometrii, ale algebry. Kluczem do rozstrzygnięcia wielkich problemów jest często odkrycie nieoczekiwanych związków między pozornie rozłącznymi obszarami matematyki. W tym konkretnym przypadku już wcześniej odkryto tego typu związek dla innych podobnych zagadnień, ale powiązanie z kwadraturą koła nie było początkowo widoczne. Nawet gdy je w końcu dostrzeżono, pojawiła się pewna trudność, której pokonanie wymagało użycia jeszcze jednej dziedziny – analizy matematycznej, czyli ścisłej wersji rachunku różniczkowego i całkowego. Na ironię zakrawa fakt, że pierwszego przełomu dokonano w czwartym dziale: w teorii liczb. Odkrycie to pozwoliło nam rozwiązać problem geometryczny, o którego rozwikłaniu starożytnym Grekom nawet się nie śniło, a prawdę mówiąc, zapewne w ogóle się nad nim nie zastanawiali – chodzi o przepis na skonstruowanie z wykorzystaniem cyrkla i linijki wielokąta foremnego o 17 bokach.

Brzmi to jak jakieś szaleństwo, szczególnie jeśli dodamy, że w taki sposób nie można skonstruować wielokątów foremnych o 7, 9, 11, 13 czy 14 bokach, natomiast jest to możliwe w wypadku figur o 3, 4,

5, 6, 8, 10 i 12 bokach. W tym szaleństwie jest jednak metoda i to właśnie ona wzbogaciła matematykę.

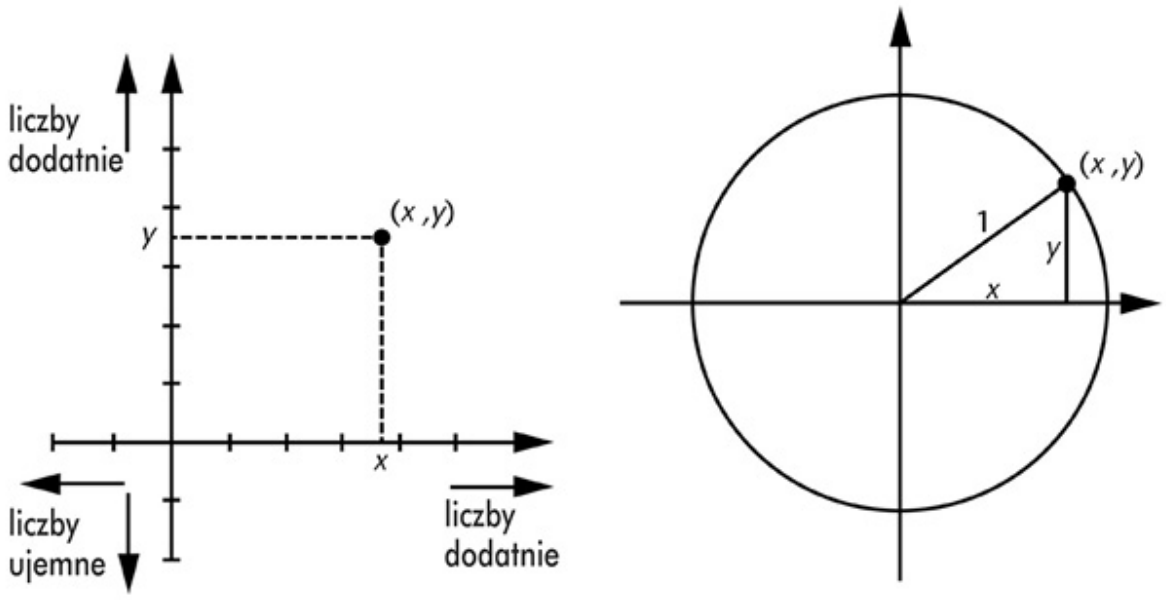
Zapytajmy jednak najpierw: czym jest wielokąt foremny? Wielokąt jest figurą geometryczną ograniczoną odcinkami. Mówimy, że figura ta jest foremna, gdy wszystkie odcinki mają taką samą długość i spotykają się pod jednakowymi kątami. Najbardziej znanym przykładem tego typu figury jest kwadrat: wszystkie jego boki mają taką samą długość, a każdy jego kąt jest kątem prostym. Istnieją również inne figury mające cztery boki jednakowej długości lub cztery równe kąty – są to, odpowiednio, romby i prostokąty. Jednak tylko kwadrat ma obie te cechy jednocześnie. Trójkąt równoboczny jest wielokątem foremnym o trzech bokach, pięciokąt foremny jest wielokątem foremnym o pięciu bokach i tak dalej (zob. ryc. 4). Euklides przedstawił opis konstrukcji klasycznych (czyli przy użyciu cyrkla i linijki) wielokątów foremnych o 3, 4 i 5 bokach. Starożytni Grecy wiedzieli również, jak można wielokrotnie podwajać liczbę boków, i tym samym potrafili skonstruować wielokąty foremne o 6, 8, 10, 12, 16, 20 bokach i tak dalej. Dzięki połączeniu ze sobą konstrukcji trójkąta i pięciokąta foremnego udało im się uzyskać przepis na wykreślenie wielokąta foremnego o 15 bokach. Na tym jednak możliwości ich metody się wyczerpały i tak już pozostało przez kolejne dwa tysiące lat. Nikt nie wyobrażał sobie, że możliwa jest konstrukcja jeszcze jakiegoś wielokąta foremnego o innej liczbie boków. Nie stawiano sobie nawet takiego pytania – wydawało się po prostu oczywiste, że w tej dziedzinie nic się już nie da zrobić.

Ryc. 4. Kilka pierwszych wielokątów foremnych. Od lewej: trójkąt równoboczny, kwadrat, pięciokąt, sześciokąt, siedmiokąt i ośmiokąt.

Potrzeba było dopiero jednego z największych matematyków w historii, żeby ktoś zastanowił się nad tym, co nie do pomyślenia, zadał pytanie, którego nikt nie ośmielił się postawić, i odkrył naprawdę zdumiewającą odpowiedź. Tym śmiałkiem był Carl Friedrich Gauss. Gauss urodził się w biednej rodzinie w Brunszwiku w Niemczech. Jego matka Dorothea nie potrafiła czytać ani pisać i nie zanotowała daty jego urodzin, pamiętała jednak, że było to w środę, osiem dni przed świętem Wniebowstąpienia w 1777 roku. Później Gauss wyznaczył dokładną datę swych urodzin z opracowanego przez siebie wzoru matematycznego do wyznaczania dat Wielkanocy. Jego ojciec Gebhard pochodził z chłopskiej rodziny, ale na życie zarabiał, wykonując różne prace dorywcze: był ogrodnikiem, rzeźnikiem ulicznym, księgowym w zakładzie pogrzebowym, a nawet kopał kanały. Ich syn był cudownym dzieckiem i, jak głosi legenda, w wieku trzech lat poprawiał już błędy rachunkowe ojca. Poznawszy się na jego nieprzeciętnym talencie do matematyki, a także do nauki języków, książę Brunszwiku wysłał go do szkoły Collegium Carolinum. Jeszcze przed jej ukończeniem Gauss niezależnie odkrył ponownie kilka ważnych twierdzeń matematycznych, udowodnionych również przez tak wybitne osobistości jak Euler. Jednak jego twierdzenie o siedemnastokącie foremnym było jak grom z jasnego nieba.

W czasach Gaussa już od 140 lat wiedziano o istnieniu bliskiego związku między geometrią i algebrą. W dodatku do *Rozprawy o metodzie* Kartezjusz sformalizował ideę, która już od jakiegoś czasu wisiała w powietrzu w niesprecyzowanej postaci – przedstawił pojęcie układu współrzędnych. Wprowadzenie układu współrzędnych sprawiło, że pusta płaszczyzna euklidesowa, zwyczajna czysta kartka papieru, stała się papierem milimetrowym pokrytym siecią kwadratów, którego inżynierowie i uczeni używają do sporządzania wykresów. Aby tego dokonać, wystarczy narysować na pustej kartce dwie proste, jedną w pionie, drugą w poziomie – linie te będziemy nazywali osiami. Teraz możemy już ustalić położenie dowolnego punktu na płaszczyźnie: wystarczy powiedzieć, jak daleko leży on od każdej z osi (zob. ryc. 5, po lewej). Te dwie liczby, które mogą

być dodatnie lub ujemne, stanowią pełny opis punktu – nazywamy je jego współrzędnymi.



Ryc. 5. Po lewej: Układ współrzędnych na płaszczyźnie. Po prawej: Wyprowadzenie równania na okrąg jednostkowy.

Wszystkie własności geometryczne punktów, prostych, okręgów i tak dalej można przełożyć na stwierdzenia algebraiczne operujące na ich współrzędnych. Bardzo trudno jest mówić z sensem o tym związku bez użycia algebry – tak jak trudno byłoby rozmawiać o piłce nożnej bez użycia słowa „gol". Na najbliższych kilku stronach pojawi się zatem kilka wzorów. Dzięki nim będziemy mogli nazwać główne postacie naszego dramatu i jednoznacznie określić łączące je związki. Znacznie prościej jest mówić „Romeo" niż „syn włoskiego szlachcica, który zakochał się w córce śmiertelnego wroga swojego ojca". Nasz Romeo będzie nosił bardziej prozaiczne imię *x*, a jego ukochaną Julią będzie *y*.

Aby zrozumieć, jak można przekształcić geometrię na zapis algebraiczny, rozważmy okrąg jednostkowy pokazany na rycinie 5 (po prawej) o środku w początku układu współrzędnych, czyli w punkcie przecięcia się osi. Zaznaczony punkt ma współrzędne (*x*,

y), a zatem dwa krótsze boki trójkąta prostokątnego pokazanego na rysunku mają długość *x* (bok poziomy) i *y* (pionowy). Najdłuższy jego bok jest promieniem okręgu, a więc ma długość 1. Z twierdzenia Pitagorasa wynika tym samym, że suma kwadratów obydwu współrzędnych jest równa 1. Używając zapisu symbolicznego, powiemy, że punkt o współrzędnych *x* i *y* leży na okręgu wtedy (i tylko wtedy), gdy jego współrzędne spełniają równanie $x^2 + y^2 = 1$. Taki symboliczny opis okręgu jest zwięzły, dokładny i świadczy o tym, że naprawdę używamy algebry. Każdą algebraiczną własność pary liczb – każde równanie wykorzystujące *x* i *y* – można zinterpretować jako stwierdzenie geometryczne mówiące o punktach, prostych, okręgach lub bardziej skomplikowanych krzywych[18].

Podstawowe równania algebraiczne wykorzystują wielomiany, czyli kombinacje potęg nieznanej wielkości *x*, w których każda potęga jest pomnożona przez pewną liczbę, zwaną współczynnikiem. Najwyższa potęga *x* pojawiająca się w równaniu jest stopniem wielomianu. Na przykład równanie

$$x^4 - 3x^3 - 3x^2 + 15x - 10 = 0$$

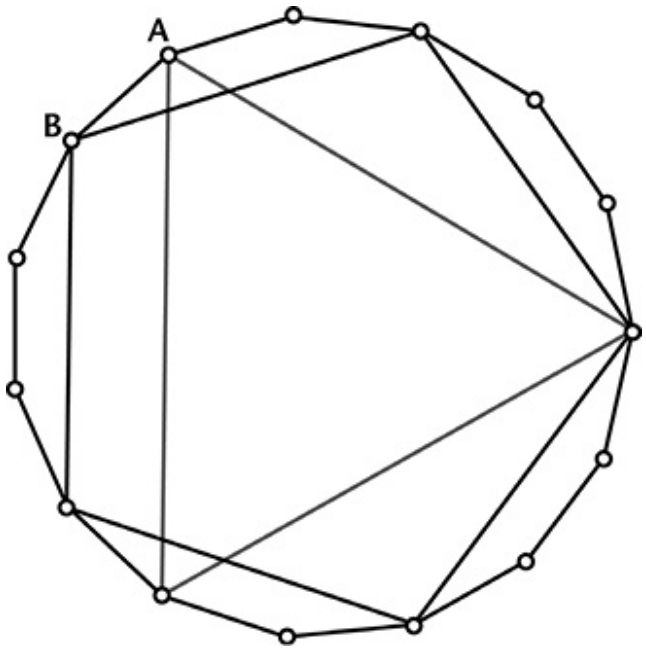
wykorzystuje wielomian rozpoczynający się od x^4, a więc wielomian stopnia czwartego. Jego współczynniki mają wartość 1, –3, –3, 15 oraz –10. Równanie to ma cztery różne rozwiązania: *x* = 1, 2, $\sqrt{5}$ i $-\sqrt{5}$ (mówimy, że liczby te są pierwiastkami tego wielomianu). Dla tych liczb lewa strona równania jest równa zero, czyli tyle samo, ile zapisano po prawej stronie znaku równości. Wielomiany stopnia pierwszego, takie jak $7x + 2$, nazywamy liniowymi i występuje w nich tylko pierwsza potęga niewiadomej. Równania stopnia drugiego, takie jak $x^2 - 3x + 2$, nazywamy równaniami kwadratowymi i występuje w nich niewiadoma w drugiej potędze – czyli podniesiona do kwadratu. W równaniu na okrąg pojawia się jeszcze druga zmienna: *y*. Jeżeli jednak znamy jeszcze jakieś drugie

równanie wiążące ze sobą *x* i *y*, na przykład równanie definiujące jakąś prostą, to możemy wyrazić *y* za pomocą *x* i zredukować równanie na okrąg do takiego, które będzie zwierało tylko *x*. To nowe równanie opisuje, jak nasza prosta przecina okrąg. W tym wypadku nowe równanie jest równaniem kwadratowym o dwóch rozwiązaniach. W ten sposób algebra odzwierciedla geometrię, w której prosta przecina okrąg w dwóch różnych punktach.

Ta cecha algebry ma duże znaczenie dla konstrukcji z użyciem cyrkla i linijki. Tego typu konstrukcje, nawet najbardziej skomplikowane, składają się z ciągu prostych kroków. W każdym kroku powstają nowe punkty w miejscach, gdzie przecinają się dwie proste, dwa okręgi lub prosta z okręgiem. Do wykreślenia tych prostych i okręgów wykorzystuje się wcześniej skonstruowane punkty. Przekładając geometrię na algebrę, można udowodnić, że równanie algebraiczne odpowiadające przecięciu się dwóch prostych jest zawsze liniowe, natomiast równanie opisujące przecięcie prostej i okręgu lub dwóch okręgów będzie kwadratowe. Wynika to z tego, że najwyższą potęgą zmiennej *x*, jaka pojawia się w równaniu okręgu, jest x^2. Zatem każdy pojedynczy krok w danej konstrukcji odpowiada rozwiązaniu równania stopnia pierwszego lub drugiego.

Bardziej złożone konstrukcje są ciągami takich podstawowych operacji i po zastosowaniu kilku technik algebraicznych można pokazać, że każda współrzędna każdego punktu skonstruowanego za pomocą cyrkla i linijki jest rozwiązaniem równania wielomianowego o całkowitych współczynnikach, którego stopień jest potęgą liczby 2. To znaczy, stopień takiego wielomianu musi być liczbą z ciągu 1, 2, 4, 8, 16 i tak dalej[19]. Jest to warunek konieczny, by dana konstrukcja mogła istnieć, można go jednak wzmocnić w taki sposób, żeby pozwalał dokładnie określić, które wielokąty foremne dają się skonstruować. W ten sposób ze skomplikowanej geometrycznej plątaniny wyłania się niespodziewanie elegancki warunek algebraiczny, który musi być spełniony dla *każdej możliwej konstrukcji*. Nie musimy nawet wiedzieć, co to jest za konstrukcja – wystarczy, że wiemy, iż opiera się na wykorzystaniu wyłącznie linijki i cyrkla.

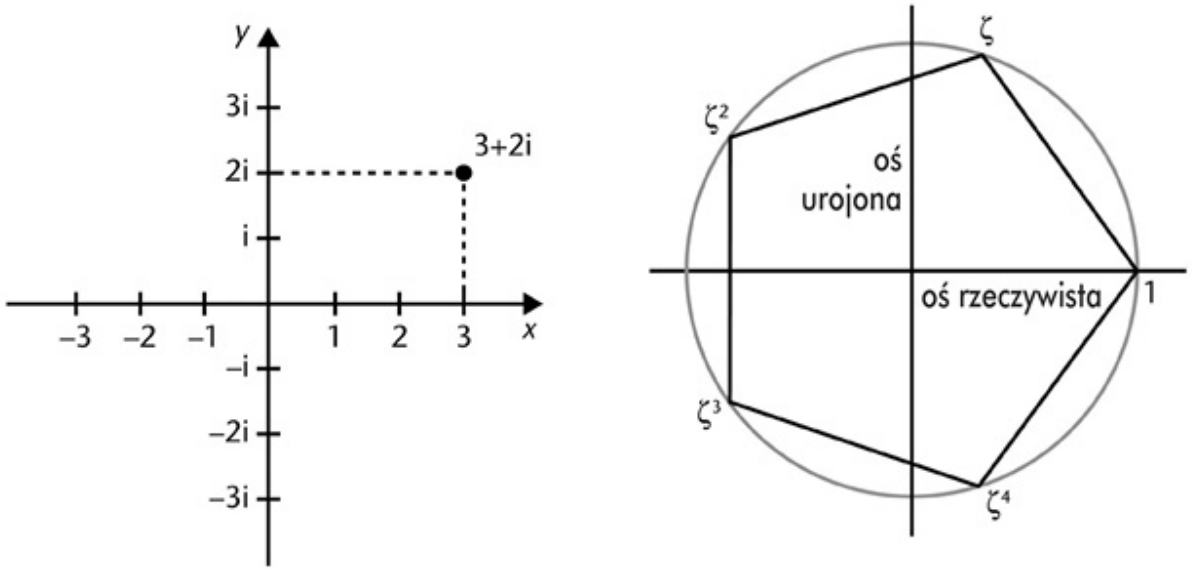
Gauss wiedział o istnieniu tej eleganckiej koncepcji. Wiedział również (ba, każdy dobrze wykształcony matematyk powinien szybko to zauważyć), że pytanie o to, jakie wielokąty foremne można skonstruować za pomocą cyrkla i linijki, sprowadza się do szczególnego przypadku wielokąta, którego liczba boków jest liczbą pierwszą. Aby zrozumieć, dlaczego tak jest, rozważmy dowolną liczbę złożoną, na przykład 15, której czynniki pierwsze to: 3 i 5 (15 = 3 × 5). Każda hipotetyczna konstrukcja wielokąta foremnego o 15 bokach (zob. ryc. 6) musi automatycznie prowadzić do uzyskania wielokąta foremnego o 3 (po uwzględnieniu co piątego wierzchołka) i 5 bokach (po uwzględnieniu co trzeciego wierzchołka). Wystarczy włożyć tylko trochę wysiłku, żeby połączyć ze sobą konstrukcje trójkąta i pięciokąta foremnego i uzyskać przepis na skonstruowanie wielokąta foremnego o 15 bokach[20]. Liczby 3 i 5 są liczbami pierwszymi i taka sama reguła obowiązuje w ogólnym przypadku. Dlatego Gauss skupił się na wielokątach, których liczba boków jest liczbą pierwszą, i postawił pytanie, jak wygląda odpowiadające im równanie. Odpowiedź jest zaskakująco elegancka. Skonstruowanie pięciokąta foremnego jest na przykład równoważne rozwiązaniu równania $x^5 - 1 = 0$. Jeśli wstawimy w miejsce liczby 5 dowolną inną liczbę pierwszą, powyższe zdanie wciąż pozostanie prawdziwe.



Ryc. 6. Konstrukcja trójkąta równobocznego i pięciokąta foremnego bazująca na konstrukcji wielokąta foremnego o 15 bokach. Aby opracować odwrotną konstrukcję, wystarczy zauważyć, że A i B są dwoma kolejnymi wierzchołkami piętnastokąta foremnego.

Wielomian ten jest stopnia piątego – a więc jego stopień *nie jest* jedną z potęg liczby 2, które wcześniej wymieniliśmy. Mimo to konstrukcja takiego wielokąta jest możliwa. Gauss szybko ustalił, dlaczego tak jest: równanie to rozpada się na dwie części, z których jedna jest stopnia pierwszego, a druga – stopnia czwartego. Zarówno 1, jak i 4 są potęgami liczby 2 i okazuje się, że kluczową rolę odgrywa tu owo równanie stopnia czwartego. Aby zrozumieć dlaczego, musimy powiązać równania z geometrią. Wymaga to użycia nowego rodzaju liczb, o których często nie wspomina się na lekcjach matematyki, ale w każdym nieco bardziej zaawansowanym zastosowaniu są one wprost nieodzowne. Chodzi o liczby zespolone, których charakterystyczną cechą jest to, że pozwalają na wyciągnięcie pierwiastka kwadratowego z minus jeden[21].

Zwyczajna liczba „rzeczywista" może być albo dodatnia, albo ujemna, ale za każdym razem jej kwadrat jest liczbą dodatnią, zatem wartość –1 nie może być kwadratem żadnej liczby rzeczywistej. Było to tak denerwujące, że matematycy wymyślili nowy rodzaj liczby, liczbę „urojoną", która po podniesieniu do kwadratu daje –1. Trzeba było oznaczyć ją jakimś symbolem, wybrano więc literę *i* (od łac. *imaginarius* – urojony). Podstawowe operacje algebraiczne – dodawanie, odejmowanie, mnożenie i dzielenie – prowadzą do powstania kombinacji liczb rzeczywistych i urojonych, na przykład: 3 + 2*i*. Liczby takie nazywamy liczbami zespolonymi, podkreślając w ten sposób fakt, że składają się z dwóch części: 3 oraz 2*i*. Liczby rzeczywiste leżą na dobrze wszystkim znanej osi liczbowej, niczym kolejne oznaczenia odległości na linijce. Liczby zespolone leżą natomiast na płaszczyźnie liczbowej, na której oś z liczbami urojonymi umieszcza się prostopadle do osi liczb rzeczywistych i dzięki temu powstaje układ współrzędnych (zob. ryc. 7, po lewej).



Ryc. 7. Po lewej: Płaszczyzna liczb zespolonych. Po prawej: Zespolone pierwiastki piątego stopnia z jedynki.

Już od dwustu lat liczby zespolone odgrywają w matematyce podstawową rolę. Obecnie matematycy uważają, że z punktu widzenia logiki należy je traktować na równi z bardziej nam znanymi liczbami „rzeczywistymi" – które przecież, tak jak wszystkie struktury matematyczne, również są pojęciami abstrakcyjnymi, a nie rzeczywistymi obiektami fizycznymi. Liczby zespolone weszły do powszechnego użycia, jeszcze zanim zainteresował się nimi Gauss, ale wciąż były uważane za zagadkowe; dopiero Gauss i kilku innych matematyków pozbawili je owej aury tajemniczości. Źródłem ich powabu jest pewien paradoks: mimo zagadkowości związanej z ich znaczeniem liczby zespolone zachowują się znacznie lepiej od liczb rzeczywistych. Zawierają brakujący składnik, którego nie można znaleźć w zbiorze liczb rzeczywistych. Pozwalają uzyskać pełny zestaw rozwiązań równań algebraicznych.

Najprostszym tego przykładem są równania kwadratowe. Jedne z nich mają dwa rzeczywiste rozwiązania, a inne – ani jednego. Na przykład równanie $x^2 - 1 = 0$ ma dwa rozwiązania: 1 i –1, natomiast równanie $x^2 + 1 = 0$ w ogóle nie ma rozwiązań rzeczywistych. Gdzieś pomiędzy nimi plasuje się równanie $x^2 = 0$, które ma tylko jedno

rozwiązanie: 0, ale w pewnym sensie jest to jedno rozwiązanie „powtórzone dwukrotnie”[22]. Gdy jednak dopuścimy rozwiązania zespolone, to okaże się, że równanie $x^2+1=0$ również ma dwa rozwiązania: *i* oraz *–i*. Gauss nie miał żadnych oporów przed wykorzystywaniem liczb zespolonych: jego praca doktorska jest pierwszym logicznie ścisłym dowodem podstawowego twierdzenia algebraicznego, że liczba zespolonych rozwiązań dowolnego równania wielomianowego jest równa (po uwzględnieniu rozwiązań wielokrotnych) stopniowi tego równania. Zatem równania kwadratowe (stopień drugi) mają zawsze dwa zespolone rozwiązania, natomiast równania sześcienne (stopień trzeci) – zawsze trzy i tak dalej.

Równanie $x^5-1=0$, które, jak powiedzieliśmy, definiuje pięciokąt foremny, jest stopnia piątego. Ma zatem pięć rozwiązań zespolonych. W zbiorze liczb rzeczywistych znajdziemy tylko jedno rozwiązanie: $x = 1$. Co z pozostałymi czterema? Opisują one cztery wierzchołki pięciokąta foremnego na płaszczyźnie zespolonej, a rozwiązanie $x = 1$ jest piątym jego wierzchołkiem (zob. ryc. 7, po prawej). Ta relacja jest przykładem matematycznego piękna: wspaniała figura geometryczna staje się eleganckim równaniem.

Zauważmy jednak, że równanie, którego rozwiązania wyznaczają owe pięć punktów, jest stopnia piątego, a 5 nie jest potęgą 2. Jednak, jak już wspomnieliśmy, nasz wielomian stopnia piątego dzieli się na dwie części o stopniach 1 i 4, które nazywamy jego nierozkładalnymi (lub nieprzywiedlnymi) czynnikami:

$$x^5-1=(x-1)\left(x^4+x^3+x^2+x+1\right)$$

(„Nierozkładalny” oznacza w tym wypadku, że taki wielomian nie ma już kolejnych czynników, przypomina więc liczby pierwsze). Pierwszy czynnik daje rozwiązanie rzeczywiste $x = 1$. Rozwiązaniem drugiego są natomiast cztery liczby zespolone wyznaczające pozostałe cztery wierzchołki pięciokąta. Jak więc widać, użycie liczb zespolonych

sprawia, że wszystko nabiera większego sensu i staje się o wiele bardziej eleganckie.

Bardzo często odtworzenie drogi, jaką działający w przeszłości matematycy doszli do jakiegoś odkrycia, jest trudne, ponieważ mieli zwyczaj publikowania tylko ostatecznego wyniku swoich przemyśleń, nie wspominając o wielu ślepych zaułkach, do których zapędzili się po drodze. Taką rekonstrukcję dodatkowo utrudnia fakt, że w przeszłości ludzie rozumowali inaczej niż obecnie. Szczególnie Gauss znany jest z tego, że zacierał za sobą wszelkie ślady poszukiwań rozwiązania i publikował tylko ostateczną analizę, wypolerowaną na wysoki połysk. Jeżeli jednak chodzi o jego badania na temat siedemnastokąta foremnego, to stąpamy po dosyć pewnym gruncie: analiza, którą ostatecznie opublikował, zawiera kilka cennych wskazówek.

Jego punkt wyjścia nie był niczym nowym. Już wcześniej kilku matematyków zwróciło uwagę na to, że omówiona przed chwilą analiza wielokątów foremnych obowiązuje również w przypadku ogólnym. Skonstruowanie wielokąta o *n* bokach jest równoważne rozwiązaniu równania $x^n - 1 = 0$ w dziedzinie liczb zespolonych. Co więcej, taki wielomian rozkłada się do postaci:

$$\left(x-1\right)\left(x^{n-1}+x^{n-2}+\cdots+x^2+x+1\right)$$

Tak jak poprzednio, pierwszy czynnik daje rozwiązanie rzeczywiste $x = 1$, a pozostałe $n - 1$ rozwiązań pochodzi z drugiego czynnika. Gdy *n* jest liczbą nieparzystą, wszystkie te rozwiązania są zespolone, natomiast gdy *n* jest parzyste, jedno z rozwiązań drugiego czynnika jest rzeczywiste: $x = -1$.

Gauss zauważył jednak coś, co umknęło wszystkim innym: czasami drugi czynnik można wyrazić za pomocą ciągu równań kwadratowych. Nie chodzi tu o przedstawienie go w postaci iloczynu prostszych czynników, ponieważ nie jest to możliwe, ale o posłużenie się równaniami, których współczynniki są rozwiązaniami innych

równań. Kluczowym faktem – naszym punktem zaczepienia przy rozwiązywaniu tego problemu – jest tu pewna elegancka własność równań algebraicznych, która staje się widoczna, gdy w taki sposób rozwiąże się ich kilka po kolei. Takie obliczenia zawsze są równoważne rozwiązaniu pojedynczego równania, ale jego stopień zazwyczaj rośnie. Zatem ceną, jaką musimy zapłacić za mniejszą liczbę równań, jest zwiększenie stopnia równania. Takie rachunki potrafią być dość zawiłe, ale jedno możemy przewidzieć: jak duży będzie stopień naszego równania. Wystarczy w tym celu pomnożyć stopnie kolejnych wielomianów.

Jeśli wszystkie równania są kwadratowe, to wynik ma postać $2 \times 2 \times \ldots \times 2$, a więc jest potęgą 2. Zatem, jeśli dana konstrukcja istnieje, to $n - 1$ musi być potęgą 2. Okazuje się jednak, że ten warunek nie zawsze wystarcza. Gdy $n = 9$, to $n - 1 = 8$, a więc jest potęgą 2. Gauss odkrył jednak, że skonstruowanie dziewięciokąta foremnego nie jest możliwe. Powodem tego jest fakt, że 9 nie jest liczbą pierwszą[23]. No dobrze, a co w przypadku kolejnej potęgi 2, która odpowiada konieczności rozwiązania ciągu czterech równań kwadratowych? Teraz stopień $n - 1$ odpowiedniego pojedynczego równania wynosi $2 \times 2 \times 2 \times 2 = 16$. Zatem $n = 17$, a to jest liczba pierwsza.

W tym momencie Gauss wiedział już, że natrafił na coś interesującego, ale musiał się uporać z jeszcze jednym problemem rachunkowym, który mógł się okazać niemożliwy do pokonania. Gauss był przekonany, że konstrukcja wielokąta foremnego o liczbie boków będącej liczbą pierwszą jest możliwa tylko wtedy, gdy liczba ta jest potęgą 2 powiększoną o 1. Uznał więc, że jest to warunek konieczny, by mogła istnieć konstrukcja – jeśli nie jest spełniony, konstrukcji takiej nie można przeprowadzić. Warunek ten jednak nie musi wcale być wystarczający – tak naprawdę istnieje mnóstwo równań stopnia 16, których nie można sprowadzić do ciągu czterech równań kwadratowych.

Jednak pewien fakt pozwalał zachować optymizm: konstrukcje przeprowadzone przez starożytnych Greków. Które liczby pierwsze się tam pojawiają? Tylko trzy: 2, 3 i 5. Wszystko to są potęgi 2

powiększone o 1, czyli: $2^0 + 1$, $2^1 + 1$ oraz $2^2 + 1$. Równanie algebraiczne związane z pięciokątem dostarcza kolejnych wskazówek. Po przemyśleniu tego wszystkiego Gauss udowodnił, że wielomian stopnia 16 związany z siedemnastokątem foremnym rzeczywiście można sprowadzić do ciągu równań kwadratowych. Konstrukcja klasyczna takiej figury musi zatem być możliwa do przeprowadzenia. Podobne podejście pozwala udowodnić, że jest tak w każdym przypadku, gdy liczba boków wielokąta jest liczbą pierwszą większą o 1 od dowolnej potęgi 2. Wnioski te świadczą o wyjątkowej zdolności Gaussa do dostrzegania matematycznych prawidłowości. Prowadzą one do pewnych ogólnych twierdzeń z zakresu teorii liczb, których nie będziemy tu omawiać. Najważniejsze w tym wszystkim jest to, że nie są to jakieś przypadkowe prawidłowości. U ich podstaw leżą solidne podwaliny. Trzeba było jednak być Gaussem, żeby je dostrzec.

Gauss nie przedstawił opisu konkretnej konstrukcji, ale opublikował wzór na rozwiązania równania 16 stopnia, które można przekształcić w taką konstrukcję, jeśli komuś na tym bardzo zależy[24]. Opisując swoje idee w *Disquisitiones Arithmeticae*, pominął wiele szczegółów, ale zapewnił, że ma pełne dowody. To ważne odkrycie sprawiło, że postanowił poświęcić życie matematyce, a nie lingwistyce. Książę w dalszym ciągu wspierał go finansowo, ale Gauss pragnął zdobyć jakieś bardziej pewne i stałe źródło dochodów. Gdy astronom Giuseppe Piazzi odkrył pierwszą asteroidę Ceres, zdołał dokonać tylko kilku obserwacji, zanim nowo odkryty świat zniknął w oślepiającym blasku Słońca. Astronomowie martwili się, że gdy asteroida w końcu odsunie się na niebie dalej od Słońca, nie uda im się jej już ponownie odnaleźć. Dokonując prawdziwego majstersztyku, który wymagał zastosowania nowych technik obliczania orbit, Gauss przewidział, gdzie asteroida powinna się ponownie pojawić – i miał rację. Dzięki temu mianowano go profesorem astronomii i dyrektorem obserwatorium w Getyndze. Zajmował to stanowisko do końca życia.

Okazuje się, że 17 nie jest jedyną nową liczbą tego rodzaju. Znamy jeszcze dwie: $2^8 + 1 = 257$ oraz $2^{16} + 1 = 65\ 537$. (Wystarczy

zastosować kilka przekształceń algebraicznych, żeby pokazać, że występująca tu liczba w wykładniku 2 sama musi być potęgą 2, w przeciwnym razie liczba taka nie może być liczbą pierwszą). Jednak dalej prawidłowość ta przestaje obowiązywać, ponieważ $2^{32} + 1 = 4\ 294\ 267\ 297$, a liczba ta jest równa iloczynowi $641 \times 6\ 700\ 417$, a zatem nie jest liczbą pierwszą. Wiadomo, że tak zwane liczby Fermata, czyli liczby postaci $2^{2^n} + 1$, są liczbami pierwszymi dla $n = 5, 6, 7, \ldots$ aż do 32. Znamy również wiele większych liczb Fermata, które nie są liczbami pierwszymi. Jak dotąd nie odkryto żadnych większych liczb pierwszych Fermata, ale w żadnym wypadku nie oznacza to, że nie mogą istnieć[25]. Potrafimy przeprowadzić konstrukcję wielokąta foremnego o 257 bokach. Pewien matematyk postawił przed sobą dość bezsensowny cel opracowania konstrukcji wielokąta o 65 537 bokach. Poświęcił na to wiele lat pracy, ale przedstawione przez niego rozwiązanie i tak zawiera błędy[25].

Z analizy Gaussa wynika, że wielokąt foremny można skonstruować za pomocą cyrkla i linijki wtedy i tylko wtedy, gdy liczba jego boków jest iloczynem jakiejś potęgi liczby 2 i *różnych* nieparzystych liczb pierwszych Fermata. W szczególności nie skonstruujemy w ten sposób dziewięciokąta foremnego. Na tej podstawie można od razu stwierdzić, że nie da się dokonać trysekcji przynajmniej jednego kąta – chodzi o kąt w trójkącie równobocznym, wynoszący 60 stopni, który po trysekcji miałby 20 stopni. Mając taki kąt, bez trudu moglibyśmy skonstruować dziewięciokąt foremny, a ponieważ jest to niemożliwe, nie można również przedstawić ogólnej konstrukcji trysekcji kąta przeprowadzonej wyłącznie za pomocą cyrkla i linijki.

Przedstawiając uzyskane wyniki, Gauss pominął wiele szczegółów swoich dowodów, ale matematycy nie mogli przecież przyjąć tych faktów na słowo honoru. W 1837 roku francuski matematyk Pierre Wantzel opublikował pełny dowód twierdzenia Gaussa o możliwości konstruowania wielokątów foremnych i przedstawił również płynący z niego wniosek, że nie można przedstawić ogólnej konstrukcji trysekcji kąta za pomocą cyrkla i linijki. Dowiódł też, że nie można skonstruować sześcianu o objętości dwukrotnie większej od objętości

sześcianu wyjściowego – to jeszcze jeden problem sięgający korzeniami starożytnej Grecji, tak zwany problem „podwojenia sześcianu".

Okazuje się, że zarówno trysekcja kąta, jak i podwojenie sześcianu są niemożliwe dlatego, że związane z tymi problemami długości są pierwiastkami nierozkładalnych wielomianów *sześciennych* – czyli wielomianów 3 stopnia. Ponieważ 3 nie jest potęgą 2, sprawa jest przesądzona. Metoda ta jednak nie sprawdza się dla problemu kwadratury koła i stoją za tym ciekawe powody. Okrąg o promieniu jednostkowym ma pole o obszarze π, a kwadrat o takim polu ma bok o długości $\sqrt{\pi}$. Istnieją metody geometryczne na wyciągnięcie pierwiastka kwadratowego, można też skonstruować kwadrat o zadanym boku, zatem rozwiązanie kwadratury koła sprowadza się do przedstawienia konstrukcji pozwalającej przekształcić odcinek o długości 1 w odcinek o długości π. Gdyby π było pierwiastkiem jakiegoś nierozkładalnego wielomianu sześciennego – lub dowolnego innego wielomianu nieredukowalnego o stopniu, który nie jest potęgą 2 – to metody Wantzela pozwoliłyby udowodnić, że przeprowadzenie kwadratury koła jest niemożliwe.

Jednak nikt nie znał żadnego równania algebraicznego, którego rozwiązaniem byłoby dokładnie π, nie mówiąc już o równaniu, którego stopień nie jest potęgą 2. Używana w szkole wartość 22/7 jest rozwiązaniem równania $7x - 22 = 0$, ale stanowi ona jedynie przybliżenie liczby π, do tego nieco za duże, informacja ta więc nie jest zbyt przydatna. Gdyby tylko udało się udowodnić, że takie równanie wielomianowe nie istnieje – a wielu matematyków podejrzewało, że tak jest, bo w przeciwnym razie już dawno byśmy je znaleźli – to niemożność rozwiązania problemu kwadratury koła byłaby oczywista. Niestety, nikt nie zdołał dowieść, że takiego równania nie ma. Liczba π tkwiła w algebraicznym stanie zawieszenia. Okazało się, że ostateczne rozwiązanie wymagało zastosowania metod wykraczających nie tylko poza geometrię – trzeba było sięgnąć po metody wykraczające nawet poza algebrę.

Aby zrozumieć, na czym polega główny problem, musimy zacząć od czegoś prostszego. W matematyce stosuje się ważne rozróżnienie

między liczbami, które można zapisać w formie dokładnych ułamków postaci p/q, gdzie *p* i *q* są liczbami całkowitymi, a takimi, których nie można tak wyrazić. Te pierwsze nazywamy liczbami wymiernymi, drugie zaś – niewymiernymi. Ułamek 22/7, będący przybliżeniem liczby π, jest na przykład liczbą wymierną. Istnieją lepsze przybliżenia: najsłynniejsze to 355/113 – wartość ta zgadza się z liczbą π z dokładnością do sześciu cyfr po przecinku. Wiadomo jednak, że żaden ułamek nie może być dokładnie równy liczbie π, ponieważ jest to liczba niewymierna. Już od bardzo dawna podejrzewano, że tak jest, ale udało się to udowodnić dopiero szwajcarskiemu matematykowi Johannowi Heinrichowi Lambertowi w 1768 roku. W swoim dowodzie wykorzystał sprytny wzór na trygonometryczną funkcję tangens, którą wyraził za pomocą ułamka łańcuchowego, czyli nieskończonego ciągu ułamków zwyczajnych[26]. W 1873 roku Charles Hermite znalazł prostszy dowód wykorzystujący wzory z dziedziny rachunku różniczkowego i całkowego. W istocie udowodnił nawet coś więcej, a mianowicie, że π^2 jest liczbą niewymierną. Zatem π nie jest nawet pierwiastkiem kwadratowym z liczby wymiernej.

Lambert podejrzewał, że prawdziwy jest znacznie silniejszy wniosek. W swoim artykule, w którym udowodnił, że π jest liczbą niewymierną, wysunął dalej idącą hipotezę, że π jest liczbą przestępną, to znaczy taką, która nie może być rozwiązaniem żadnego równania wielomianowego o współczynnikach całkowitych. Późniejsze odkrycia pokazały, że miał rację. Przełom nastąpił w dwóch etapach. Nowa metoda Hermite'a zastosowana w dowodzie na niewymierność przygotowała grunt pod dalsze prace, pokazując, że dobrą strategią może być tu użycie rachunku różniczkowego i całkowego – a mówiąc dokładnie, jego ścisłej odmiany, czyli analizy matematycznej. Idąc dalej tym tropem, Hermite odkrył wspaniały dowód, że inna interesująca liczba pojawiająca się często w matematyce, a mianowicie podstawa logarytmu naturalnego *e*, jest liczbą przestępną. Wartość *e* jest równa w przybliżeniu 2,71828 i można powiedzieć, że jest to liczba jeszcze ważniejsza niż π. Przedstawiony przez Hermite'a dowód przestępności tej liczby jest

wprost magiczny, niczym królik wyciągnięty teatralnym gestem z cylindra analizy matematycznej. Owym królikiem jest skomplikowany wzór związany z hipotetycznym równaniem algebraicznym. Hermite założył na początku dowodu, że liczba *e* jest rozwiązaniem tego równania, a następnie posłużył się algebrą, żeby pokazać, że wzór ten musi być równy niezerowej liczbie całkowitej. W ostatnim kroku sięgnął do analizy matematycznej i dowiódł, że liczba ta musi leżeć w przedziale między $-\frac{1}{2}$ i $\frac{1}{2}$. Ponieważ jedyną liczbą całkowitą w tym przedziale jest zero, doszedł w ten sposób do sprzeczności. Tym samym założenie, że *e* jest rozwiązaniem równania algebraicznego, musi być fałszywe, a więc *e* jest liczbą przestępną.

W 1882 roku Ferdinand Lindemann dodał parę wodotrysków do metody Hermite'a i dowiódł, że jeśli jakaś liczba niezerowa jest rozwiązaniem równania algebraicznego, to *e* podniesione do potęgi równej tej liczbie nie może być rozwiązaniem równania algebraicznego. W dowodzie wykorzystał znany już Eulerowi związek między liczbami π i *e* oraz jednostką urojoną *i* – słynny wzór $e^{i\pi} = -1$. Załóżmy, że π jest rozwiązaniem pewnego równania algebraicznego. W takim wypadku również *iπ* musi być rozwiązaniem jakiegoś równania i z twierdzenia Lindemanna wynika wówczas, że –1 *nie może* być rozwiązaniem równania algebraicznego. To jednak w sposób oczywisty jest niezgodne z prawdą: liczba –1 jest rozwiązaniem równania $x+1=0$. Jedynym wyjściem z tej sprzeczności jest przyjęcie, że π nie jest rozwiązaniem równania algebraicznego, a więc że jest liczbą przestępną. A to oznacza, że nie można przeprowadzić konstrukcji kwadratury koła.

Przebyliśmy długą i okrężną drogę od geometrii Euklidesa do dowodu Lindemanna, zajęło nam to ponad dwa tysiące lat, ale w końcu tam dotarliśmy. Nie jest to jednak opowieść jedynie o tym, że nie można wykonać kwadratury koła. To prawdziwa lekcja poglądowa pokazująca, jak należy rozwiązywać wielkie problemy matematyczne. Matematycy musieli najpierw określić precyzyjnie, co rozumieją przez pojęcie „konstrukcji geometrycznej". Musieli opisać

ogólne cechy takich konstrukcji, ponieważ z tego wynikają ograniczenia na to, co można za ich pomocą osiągnąć. Znalezienie tych cech wymagało powiązania geometrii z innym działem matematyki – z algebrą. Rozwiązywanie zadań algebraicznych, nawet w najprostszych przypadkach, takich jak konstrukcja wielokątów foremnych, wymaga z kolei zastosowania teorii liczb. Uporanie się z trudnym przypadkiem liczby π wymagało zastosowania nowego podejścia i rozwiązania trzeba było szukać w jeszcze innej dziedzinie – analizie matematycznej.

Żaden z tych kroków nie był prosty ani oczywisty. Dokończenie dowodu zajęło uczonym około stu lat, nawet wtedy, gdy wszystkie główne idee były już na miejscu. Rozwiązaniem tych problemów zajęli się najlepsi matematycy epoki, a co najmniej jeden z nich był najlepszym matematykiem *wszech czasów*. Do rozwiązania wielkich problemów potrzebne jest dogłębne zrozumienie matematyki, upór w dążeniu do celu i pomysłowość. W prace nad tak trudnymi zadaniami trzeba włożyć wiele lat wytężonego wysiłku i liczyć się z tym, że większość trudu może pójść na marne. Wyobraźmy sobie jednak, jakie to musi być uczucie, gdy w końcu upór przynosi owoce i udaje się wyjaśnić coś, co dręczyło ludzkość przez całe stulecia. Warto w tym miejscu przytoczyć słowa prezydenta Johna F. Kennedy'ego wypowiedziane w 1962 roku, gdy ogłosił, że Stany Zjednoczone zamierzają wysłać człowieka na Księżyc: „Postawiliśmy sobie taki cel nie dlatego, że to łatwe, ale właśnie dlatego, że jest to trudne".

Większość opowieści matematycznych nigdy się tak naprawdę nie kończy i historia liczby π nie jest pod tym względem wyjątkiem. W 1997 roku Fabrice Bellard ogłosił[27], że bilionową cyfrą liczby π w zapisie binarnym jest 1. Była to niezwykła wiadomość, ale nie ze względu na przedstawioną odpowiedź. Zdumiewające było to, że Bellard nie wyliczył żadnej z wcześniejszych cyfr. Po prostu wybrał sobie na chybił trafił jedną z cyfr binarnego rozwinięcia liczby π.

Obliczenie jej wartości było możliwe dzięki zastosowaniu dziwnego wzoru na liczbę π odkrytego w 1996 roku przez Davida Baileya,

Petera Borweina i Simona Plouffe'a. Może się on wydawać nieco skomplikowany, ale warto mu się uważnie przyjrzeć:

$$\pi = \sum_{n=0}^{\infty} \frac{1}{2^{4n}} \left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6} \right)$$

Wielka litera Σ oznacza sumę podanych wyrazów w określonym przedziale. W naszym przykładzie *n* zmienia się od 0 do nieskończoności (∞). Tak naprawdę Bellard posłużył się wzorem, który sam wyprowadził, wykorzystując podobne metody. Jego wzór jest nieco lepszy z punktu widzenia szybkości przeprowadzania obliczeń:

$$\pi = \frac{1}{64} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{10n}} \left(-\frac{32}{4n+1} - \frac{1}{4n+3} + \frac{256}{10n+1} - \frac{64}{10n+3} - \frac{4}{10n+5} - \frac{4}{10n+7} + \frac{1}{10n+9} \right)$$

Kluczowym faktem jest to, że wiele z pojawiających się w tym wzorze liczb – 1, 4, 32, 64, 256, a także 2^{4n} i 2^{10n} – to potęgi 2, które można oczywiście bardzo prosto przedstawić w zapisie binarnym używanym wewnętrznie w komputerach. Odkrycie to zapoczątkowało całą lawinę nowych wzorów na wartość liczby π, a także kilku innych interesujących liczb. Obecnie regularnie już co jakiś czas dowiadujemy się o pobiciu rekordu Bellarda i obliczeniu jeszcze dalszej cyfry binarnego rozwinięcia liczby π: w 2010 roku Nicholas Sze z firmy Yahoo wyznaczył cyfrę tego rozwinięcia znajdującą się na dwubiliardowej pozycji – okazało się, że jest to 0.

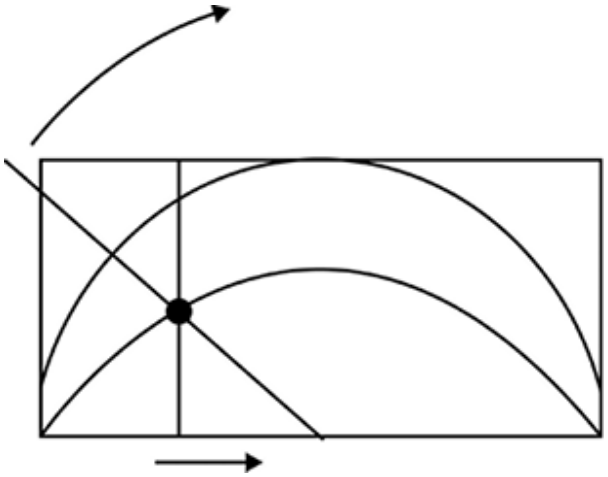
Używając tego samego wzoru, można wyznaczyć pojedyncze cyfry liczby π w systemach liczbowych o podstawie 4, 8 i 16. Nie znamy podobnego wzoru dla żadnych innych systemów liczbowych. W szczególności nie potrafimy wyznaczać pojedynczych cyfr w zapisie dziesiętnym. Czy takie wzory istnieją? Przed pojawieniem się wzoru Baileya, Borweina i Plouffe'a nikt sobie nie wyobrażał, że można czegoś takiego dokonać w systemie dwójkowym.

14 Szczegółowe informacje na ten temat (w języku angielskim) można znaleźć na stronie internetowej: http://www.numberworld.org/misc_runs/pi-5t/details.html

15 Problem znalezienia skończonego podziału koła na części i ułożenia ich następnie w taki sposób, żeby utworzyły kwadrat o takim samym polu, nosi nazwę problemu kwadratury koła Tarskiego. Rozwiązał go Miklós Laczkovich w 1990 roku. Zastosowana przez niego metoda jest dowodem niekonstruktywnym i opiera się na wykorzystaniu aksjomatu wyboru. Metoda ta wymaga podziału koła na olbrzymią liczbę fragmentów, a mianowicie na 10^{50} części.

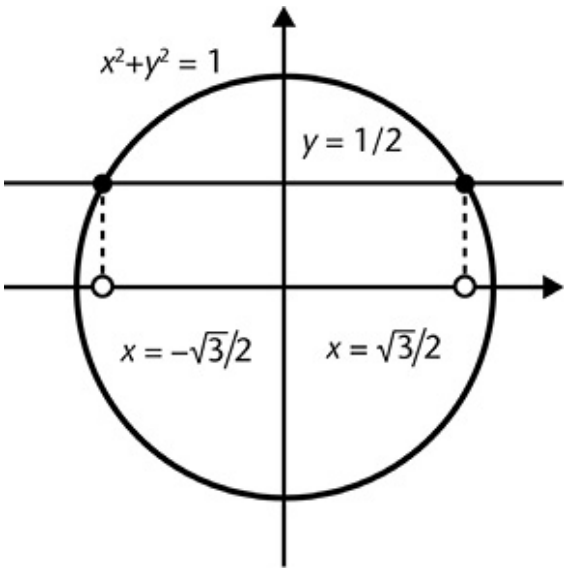
16 Underwood Dudley przeanalizował dogłębnie takie dziwaczne „dowody" kwadratury koła i trysekcji kąta w swojej książce *A Budget of Trisections* (*Zapas trysekcji*). Nie jest to nowe zjawisko – zob. Agustus De Morgan, *A Budget of Paradoxes* (*Zapas paradoksów*), Longmans, 1872; przedruk: Books For Libraries Press, 1915.

17 Kwadratrysa Hippiasza z Elidy jest krzywą kreśloną przez punkt przecięcia linii pionowej przesuwającej się jednostajnie w poprzek prostokąta z linią, która obraca się ze stałą prędkością kątową wokół środka jego dolnego boku (zob. ryc. 52). Krzywa ta pozwala przekształcić każde pytanie na temat podziału kąta w odpowiadające mu pytanie o podział odcinka. Na przykład aby dokonać trysekcji kąta, wystarczy podzielić odpowiedni odcinek na trzy równe części. Więcej informacji w języku angielskim można znaleźć na stronie internetowej http://www.geom.uiuc.edu/~huberty/math5337/groupe/quadratrix.html.



Ryc. 52. Kwadratrysa Hippiasza (dolna krzywa)

18 Oto konkretny przykład: z geometrii wiemy, że prosta przecinająca okrąg, która nie jest jego styczną, ma z nim dokładnie dwa punkty wspólne. Rozważmy prostą równoległą do poziomej osi układu współrzędnych, przebiegającą w odległości $\frac{1}{2}$ od niej (zob. ryc. 53). Równanie takiej prostej jest nieskomplikowane: $y=\frac{1}{2}$. (Bez względu na to, jaką wartość przyjmuje *x*, zawsze dostajemy tę samą wartość zmiennej *y*). Dla $y=\frac{1}{2}$ równanie $x^2+y^2=1$ przyjmuje postać $x^2+\frac{1}{4}=1$. Zatem $x^2=\frac{3}{4}$, a więc $x=\frac{\sqrt{3}}{2}$ lub $-\frac{\sqrt{3}}{2}$. W ten sposób algebra mówi nam, że okrąg jednostkowy przecina naszą prostą dokładnie w dwóch punktach o współrzędnych: $\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)$ oraz $\left(-\frac{\sqrt{3}}{2},\frac{1}{2}\right)$. Zgadza się to z tym, co możemy odczytać z ryciny 53, analizując konstrukcję czysto geometryczną.



Ryc. 53. Linia pozioma przecina okrąg w dwóch punktach.

19 Mówiąc ściślej, taki wielomian musi mieć całkowite współczynniki i nie może być rozkładalny, to znaczy nie może być iloczynem dwóch wielomianów niższego stopnia o całkowitych współczynnikach. Sam fakt, że wielomian ma stopień będący potęgą 2, nie zawsze wystarcza, by mogła istnieć konstrukcja klasyczna, ale zawsze jest to konieczne. Jeśli stopień wielomianu nie jest potęgą 2, nie ma mowy o możliwości konstrukcji z użyciem cyrkla i linijki. Jeśli jest potęgą 2, to należy przeprowadzić dalszą analizę, by stwierdzić, czy konstrukcja klasyczna istnieje

20 Prawdą jest również odwrotne stwierdzenie: jeśli potrafimy skonstruować trójkąt i pięciokąt foremny, to z tych konstrukcji możemy wyprowadzić konstrukcję piętnastokąta foremnego. Wynika to z faktu, że $2/5 - 1/3 = 1/15$. Szczególnej uwagi wymagają jednak potęgi liczb pierwszych. Taka argumentacja nie pozwala na wyprowadzenie konstrukcji, powiedzmy, dziewięciokąta na podstawie znajomości konstrukcji figury odpowiadającej jednemu z czynników pierwszych liczby 9 – a więc trójkąta. Gauss udowodnił, że tego typu konstrukcja nie jest możliwa dla nieparzystych liczb pierwszych występujących w potędze większej niż jeden.

21 Ian Stewart, *17 równań, które zmieniły świat*, przeł. Julia Szajkowska, Prószyński i S-ka, Warszawa 2013, rozdział 5.

22 Aby zrozumieć, skąd bierze się taka interpretacja, wystarczy rozłożyć wielomian kwadratowy na czynniki liniowe. Na przykład wielomian $x^2 - 1 = (x+1)(x-1)$ przyjmuje wartość zero, gdy dowolny z tych czynników staje się zerem, a więc gdy $x = 1$ lub -1. To samo rozumowanie można zastosować dla $x^2 = xx$: ten wielomian przyjmuje wartość zero, gdy pierwszy lub drugi czynnik $x = 0$. Akurat w tym przypadku oba rozwiązania dają taką samą wartość x, ale to wystąpienie dwóch czynników x odróżnia tę sytuację od wielomianu na przykład postaci $x(x-1)$, w którym mamy tylko jeden czynnik x. Gdy musimy znaleźć liczbę rozwiązań równania algebraicznego, odpowiedź nastręcza zwykle mniejszych problemów, jeśli uwzględnimy takie „wielokrotności”.

23 Gdy $n = 9$, to drugi czynnik ma postać:

$$x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$$

Jednak ten wielomian rozkłada się na dwa czynniki – jest on równy:

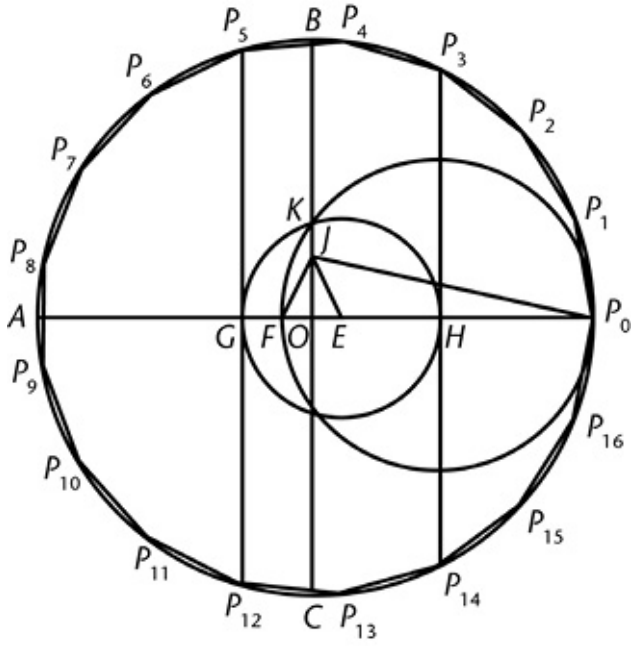
$$(x^2 + x + 1)(x^6 + x^3 + 1)$$

Z analizy Gaussa wynika, że dana konstrukcja jest możliwa do przeprowadzenia tylko wtedy, gdy każdy nieprzywiedlny czynnik ma stopień będący potęgą 2. W tym jednak wypadku drugi czynnik jest wielomianem stopnia 6, a 6 nie jest potęgą 2.

24 Gauss udowodnił, iż konstrukcja siedemnastokąta foremnego jest możliwa, pod warunkiem że potrafimy skonstruować odcinek o długości

$$\frac{1}{16}\left[-1+\sqrt{17}+\sqrt{34-2\sqrt{17}}+\sqrt{68+12\sqrt{17}-16\sqrt{34+2\sqrt{17}}-2\left(1-\sqrt{17}\right)\left(\sqrt{34-2\sqrt{17}}\right)}\right]$$

Ponieważ konstrukcja pierwiastków kwadratowych jest zawsze możliwa, problem jest w zasadzie rozwiązany. Inni matematycy przedstawili dokładne opisy tego typu konstrukcji. Pierwszą opublikował Ulrich von Huguenin w 1803 roku, a w 1893 roku H.W. Richmond przedstawił prostszą konstrukcję. Przyjmując oznaczenia takie jak na rycinie 54, zaznaczmy dwie prostopadłe średnice okręgu: AOP_0 oraz BOC. Wyznaczmy punkt J taki, że OJ = $^1/_4$OB, a kąt OJE = $^1/_4$OJP_0. Znajdźmy punkt F taki, że kąt EJF wynosi 45 stopni. Narysujmy okrąg, którego średnicą jest FP_0 przecinający odcinek OB w punkcie K. Narysujmy okrąg o środku w E i przechodzący przez K, który przecina odcinek AP_0 w punktach G i H. Narysujmy odcinki HP_3 i GP_5 prostopadłe do AP_0. W ten sposób wyznaczyliśmy punkty P_0, P_3 oraz P_5, które są, odpowiednio, zerowym, trzecim i piątym wierzchołkiem siedemnastokąta foremnego. Pozostałe wierzchołki można wyznaczyć już w prosty sposób.



Ryc. 54. Konstrukcja siedemnastokąta foremnego

25 Friedrich Julius Richelot przedstawił opis konstrukcji 257-kąta foremnego w 1832 roku. Johann Gustav Hermes z uniwersytetu w Lingen poświęcił dziesięć lat na opracowanie konstrukcji 65537-kąta foremnego. Jego nieopublikowaną pracę można znaleźć na uniwersytecie w Getyndze, ale uważa się, że zawiera ona błędy.

26 Najnowsze informacje na temat badań takich liczb można znaleźć na stronie internetowej Wilfrida Kellera *Prime factors of Fermat numbers and complete factoring status* (*Czynniki pierwsze liczb Fermata i pełna informacja o postępach prac*): http://www.prothsearch.net/fermat.html.

Typowy ułamek łańcuchowy wygląda mniej więcej tak:

$$3+\cfrac{1}{7+\cfrac{1}{15+\cfrac{1}{1+\cfrac{1}{292+\cdots}}}}$$

Ten konkretny ułamek łańcuchowy jest początkiem ułamka odpowiadającego liczbie π.

27 Zob.: http://bellard.org/pi-challenge/announce220997.html.

4. Tajemnice kartografii

Twierdzenie o czterech barwach

Bardzo często wielkie problemy matematyczne wynikają z niemożności udzielenia odpowiedzi na głębokie i trudne pytania dotyczące ugruntowanych już dziedzin matematyki. Są wielkimi wyzwaniami, które pojawiają się, gdy jakiś ważny obszar zostaje dokładnie zbadany. Zwykle ich zrozumienie wymaga specjalistycznej wiedzy i wszyscy uczeni zajmujący się daną dziedziną wiedzą, że znalezienie odpowiedzi na te pytania jest trudne, ponieważ wielu już próbowało, ale nikomu się nie udało. Dziedzina taka ma najczęściej bogaty arsenał różnych technik obliczeniowych, potężne matematyczne działa, którymi może się posługiwać każdy, kto przyswoił sobie odpowiednią wiedzę – skoro jednak dany problem wciąż nie został rozwiązany, oznacza to, że wszystkie możliwe sposoby zastosowania tych technik zostały już wypróbowane i okazało się, że w tym wypadku *nie działają*. Zatem albo musi istnieć jakiś nieszablonowy sposób zastosowania wypróbowanych technik z tej dziedziny, albo trzeba wymyślić nową metodę.

Historia matematyki pokazuje, że każde z tych podejść może przynieść owoce.

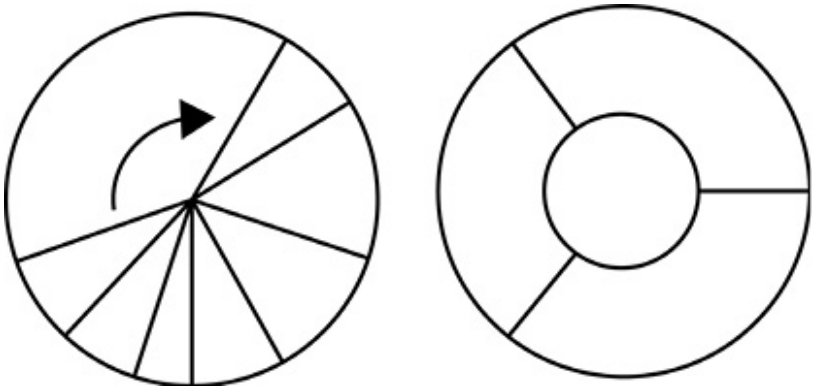
Zdarzają się jednak również wielkie problemy, które zupełnie nie pasują do tego opisu. Pojawiają się znikąd – gryzmoły na piasku, notatka na marginesie, czyjaś przelotna myśl. Można je sformułować w prosty sposób, ale ponieważ nie stoi za nimi żadna rozbudowana dziedzina matematyki, nie mamy ustalonych metod ich rozwiązywania. Mogą upłynąć całe lata, zanim ktoś dostrzeże, jak są w istocie trudne: nie sposób przecież wykluczyć, że może istnieć

jakaś sprytna, ale prosta sztuczka pozwalająca rozwiązać je w kilku linijkach. Problem czterech barw należy do tej drugiej kategorii. Upłynęły całe dziesięciolecia, zanim matematycy zaczęli zdawać sobie sprawę, jak trudne jest to wyzwanie, a na domiar złego przez większość tego czasu sądzili, że problem *został* już rozwiązany na kilku zaledwie stronach. Wydawało się, że jest to problem z pogranicza nauki, mało kto więc traktował go poważnie. A gdy już się ktoś taki znalazł, okazywało się, że przedstawione przez niego rozwiązanie jest niedoskonałe. Ostatecznie udało się przedstawić rozwiązanie usuwające wszystkie te niedoskonałości, ale wówczas cały wywód tak się skomplikował, że uczeni musieli skorzystać z pomocy potężnych komputerów.

Po dłuższym czasie oba rodzaje problemów upodobniają się do siebie, ponieważ mimo różnej genezy rozwiązanie ich wymaga w obu wypadkach zastosowania nowych sposobów myślenia. Wprawdzie problemy pierwszego rodzaju są osadzone w dobrze poznanych obszarach, ale tradycyjne metody nie zdają w odniesieniu do nich egzaminu. Problemy drugiego rodzaju nie należą do żadnego znanego obszaru – w istocie, często prowadzą do powstania nowych dziedzin – nie ma więc żadnych ugruntowanych technik, które można by zastosować do ich rozwiązania. W obu wypadkach rozwiązanie wymaga wynalezienia nowych metod i odkrycia nowych związków z innymi działami matematyki.

Wiemy bardzo dobrze, skąd wziął się problem czterech barw – i wcale nie pojawił się na gruncie matematyki. W 1852 roku Francis Guthrie, młody matematyk i botanik z Afryki Południowej, kończący właśnie studia prawnicze, postanowił pokolorować hrabstwa na mapie Anglii. Chciał to zrobić w taki sposób, żeby wszystkie sąsiadujące ze sobą hrabstwa były oznaczone różnymi kolorami, dzięki czemu ich granice byłyby wyraźnie widoczne. Guthrie odkrył, że wystarczają do tego tylko cztery barwy. Później dokonał jeszcze kilku innych prób tego rodzaju i doszedł do wniosku, że cztery barwy powinny wystarczyć do pokolorowania w ten sposób dowolnej mapy. Za „sąsiednie” uważał przy tym obszary mające wspólną granicę

o niezerowej długości – jeśli dwa hrabstwa stykały się tylko w jednym punkcie lub w kilku oddzielnych punktach, można je było pokolorować tak samo. Bez takiego zastrzeżenia nie można nałożyć żadnego ograniczenia na liczbę potrzebnych kolorów, ponieważ w jednym punkcie może się stykać dowolna liczba obszarów (zob. ryc. 8, po lewej).



Ryc. 8. Po lewej: W pojedynczym punkcie może się zbiegać dowolna liczba obszarów. Po prawej: Potrzebne są przynajmniej cztery kolory.

Guthriego zainteresowało, czy jego odkrycie jest znanym twierdzeniem matematycznym, spytał więc o to swojego brata Fredericka, który studiował matematykę pod kierunkiem wybitnego, choć ekscentrycznego Augustusa De Morgana na University College w Londynie. De Morganowi nie było nic na ten temat wiadomo, napisał więc do jeszcze wybitniejszego matematyka, Irlandczyka Williama Rowana Hamiltona:

> Jeden z moich studentów [później ustalono, że był to Frederick Guthrie] poprosił mnie dzisiaj, żebym mu przedstawił uzasadnienie pewnego faktu, o którego istnieniu nie miałem pojęcia – i wciąż nie jestem przekonany, czy rzeczywiście jest to fakt. Twierdzi on mianowicie, że jeśli jakąś figurę podzieli się w dowolny sposób, a następnie pokoloruje się powstałe figury różnymi barwami tak, żeby dwa obszary oddzielone

> wspólną *linią* graniczną miały różny kolor – to działanie takie może wymagać użycia czterech barw, ale nie więcej. […] Zastanawiam się, czy nie można przedstawić przykładu wymagającego użycia pięciu lub więcej barw. […] Co pan o tym sądzi? I, jeśli to prawda, czy ktoś już zwrócił na to uwagę?

Frederick wspominał później o „dowodzie" zaproponowanym przez jego brata, ale twierdził również, że podstawową ideą tego dowodu jest rysunek taki jak pokazany na rycinie 8, z którego wynika jedynie, iż trzeba użyć co najmniej czterech kolorów.

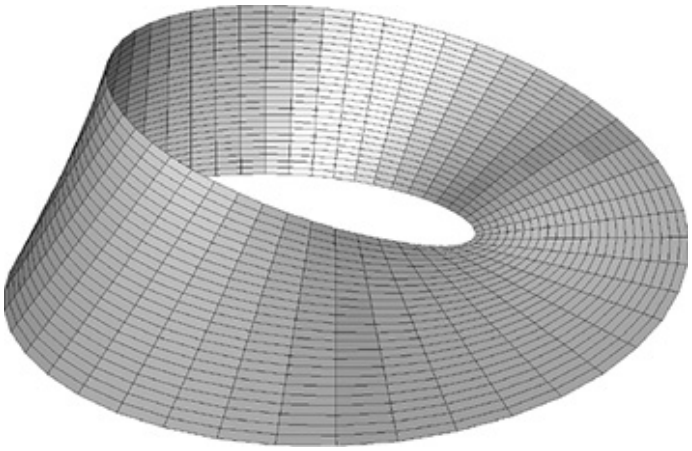
Odpowiedź Hamiltona była krótka i niezbyt pomocna: „W najbliższym czasie nie będę mógł się raczej zająć Pańskim »kwaternionem« kolorów" – odpisał. W tym okresie pracował już nad systemem algebraicznym, któremu poświęcił całe swoje życie. Jest to struktura algebraiczna podobna do liczb zespolonych, ale obejmująca cztery rodzaje liczb zamiast dwóch, jak to jest w przypadku liczb zespolonych. Nazwał te liczby „kwaternionami". W dalszym ciągu odgrywają one istotną rolę w matematyce, a można nawet powiedzieć, że teraz stały się jeszcze ważniejsze, niż były w czasach Hamiltona. Nigdy jednak nie nabrały tak dużego znaczenia, jakiego życzyłby sobie Hamilton. Użycie w liście określenia „kwaternion" było jedynie naukowym żartem i długo wydawało się, że nie ma żadnego związku między kwaternionami a problemem czterech barw. Istnieje jednak pewne sformułowanie tego problemu, które można rozumieć jako stwierdzenie dotyczące kwaternionów, a więc dowcip Hamiltona miał drugie dno[28].

Mimo prób De Morgan nie potrafił znaleźć dowodu i wspomniał o tym problemie znajomym matematykom w nadziei, że któryś z nich podsunie jakieś rozwiązanie. Pod koniec lat sześćdziesiątych XIX wieku amerykański logik, matematyk i filozof Charles Sanders Peirce oznajmił, że rozwiązał problem czterech barw, a także że znalazł odpowiedzi na podobne pytania na temat map na bardziej skomplikowanych powierzchniach. Jego rzekomy dowód nigdy nie

został opublikowany i wątpliwe jest, żeby dostępne wówczas metody były do tego celu wystarczające.

Chociaż w sformułowaniu problemu czterech barw mówi się o mapach, nie ma on większego znaczenia dla kartografii. Względy praktyczne brane pod uwagę przy kolorowaniu map odzwierciedlają głównie różnice w poglądach politycznych i jeśli to oznacza, że dwa sąsiednie obszary muszą mieć taki sam kolor, to trudno, niech tak będzie. Problem ten okazał się jednak interesujący na gruncie matematyki teoretycznej, w nowym dziale badań, który zaczął się dopiero wykształcać, a mianowicie w topologii. Chodzi o słynną „gumową geometrię", w której figury można odkształcać w dowolny ciągły sposób. Nawet jednak w topologii problem czterech barw nie miał związku z głównym nurtem badań. Wydawało się, że nie jest to nich ważnego, ot, taka sobie ciekawostka.

Jednym z pionierów topologii był August Möbius, znany obecnie głównie za sprawą jednopowierzchniowej wstęgi nazwanej jego imieniem (zob. ryc. 9). Nietrudno sporządzić jej model. Wystarczy wziąć pasek papieru, zwinąć go w kształt pierścienia, obrócić jeden koniec o 180 stopni i skleić oba końce. Kiedyś znajomy Möbiusa, lingwista Benjamin Weiske, zadał mu zagadkę: czy indyjski władca mający pięciu synów może podzielić między nich swoje królestwo w taki sposób, żeby każda prowincja miała niezerową granicę z prowincjami pozostałych czterech książąt? Möbius powtórzył tę zagadkę swoim studentom w ramach ćwiczenia. Na następnym wykładzie przeprosił ich jednak, że dał im zadanie niemożliwe do rozwiązania. Chciał przez to powiedzieć, że udało mu się *udowodnić*, iż jest to niemożliwe[29].

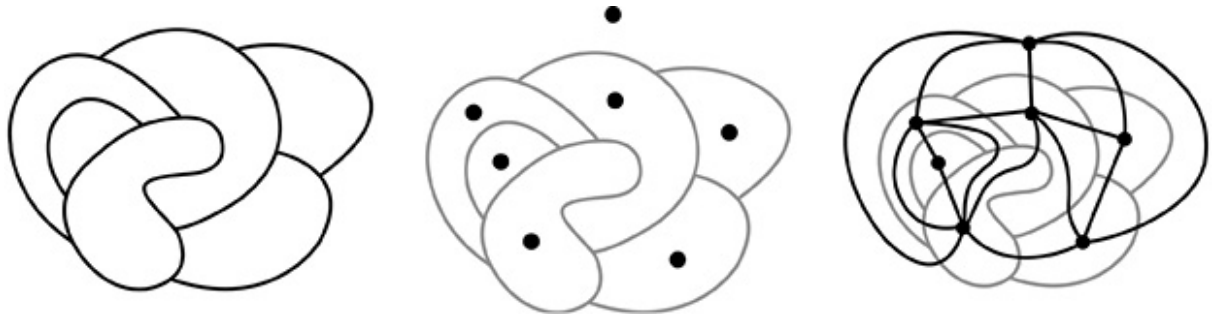


Ryc. 9. Wstęga Möbiusa ma tylko jedną stronę.

Trudno byłoby rozwiązać tę zagadkę na gruncie geometrii, ponieważ kształt poszczególnych obszarów i ich wzajemny układ mógłby być w zasadzie bardzo skomplikowany. Postępów dokonamy jedynie wtedy, gdy wprowadzimy duże uproszczenie: tak naprawdę znaczenie ma tylko to, które obszary sąsiadują ze sobą i jak są ułożone względem siebie ich granice. Jest to informacja o charakterze topologicznym, nie zależy bowiem od konkretnych kształtów. Można ją przedstawić w jednoznaczny, prosty sposób za pomocą diagramu nazywanego grafem lub siecią – ta druga nazwa jest obecnie częściej używana, wywołuje bowiem silniejsze skojarzenia.

Sieć jest zabójczo prostym pojęciem: jest to zbiór węzłów, rysowanych jako kropki, które mogą być połączone krawędziami, przedstawianymi na rysunku za pomocą linii. Weźmy dowolną mapę, na przykład tę z ryciny 10 (po lewej). Aby przekształcić ją w sieć, należy narysować kropkę wewnątrz każdego obszaru (zob. ryc. 10, w środku). Jeśli dwa obszary mają wspólną granicę, musimy narysować linię łączącą odpowiednie punkty, przechodzącą przez określony odcinek granicy. Jeżeli dane obszary mają kilka oddzielnych odcinków granicznych, każdy z nich musi otrzymać własną linię. Należy w ten sposób przekształcić wszystkie obszary i wszystkie odcinki granic, ale tak, żeby linie nie przecinały się nawzajem ani same ze sobą i zbiegały jedynie w narysowanych

wcześniej punktach. Później możemy już wyrzucić mapę i zachować tylko kropki i linie. W ten sposób otrzymaliśmy sieć dualną naszej mapy (zob. ryc. 10, po prawej)[30].



Ryc. 10. Po lewej: Mapa. W środku: Należy umieścić punkt w każdym obszarze. Po prawej: Po połączeniu punktów liniami przecinającymi granice otrzymujemy sieć dualną (tworzą ją tylko punkty i czarne linie).

Użyliśmy tu słowa „dualna", ponieważ ta procedura zastępuje obszary, linie i punkty (miejsca, w których zbiegają się poszczególne obszary) punktami, liniami i obszarami. Każdemu obszarowi na mapie odpowiada punkt w sieci dualnej. Każdemu odcinkowi granicy odpowiada linia w sieci dualnej – nie jest to jednak taka sama linia jak granica na mapie, ponieważ krawędź sieci przecina granicę i łączy ze sobą dwa wierzchołki. Każdemu punktowi na mapie, w którym styka się trzy lub więcej obszarów, odpowiada obszar w sieci dualnej ograniczony pętlą utworzoną z krawędzi. Zatem sieć dualna sama jest mapą, ponieważ jej linie ograniczają pewne obszary. Okazuje się też, że siecią dualną sieci dualnej jest mapa wyjściowa, z dokładnością do kilku szczegółów polegających na pominięciu niepotrzebnych punktów i linii.

Możemy teraz sformułować zagadkę o pięciu książętach, wykorzystując pojęcie sieci dualnej: czy można na płaszczyźnie połączyć ze sobą pięć punktów liniami tak, by się nie przecinały? Odpowiedź brzmi: „nie", a kluczem do niej jest wzór Eulera, który stwierdza, że jeśli mapa na płaszczyźnie składa się z *S* ścian

(obszarów), *K* krawędzi (linii) i *W* wierzchołków (punktów), to $S + W - K = 2$. W naszym przykładzie całą resztę płaszczyzny, czyli wszystko, co jest na zewnątrz sieci, należy potraktować jako jeden wielki obszar. Wzór ten był pierwszą wskazówką, że rozważania topologiczne nie muszą być stratą czasu – wrócimy do niego w rozdziale 10.

Dowód na to, że rozwiązanie zagadki o indyjskich książętach jest niemożliwe, zaczyna się od założenia, że takie rozwiązanie istnieje, a następnie w kolejnych krokach dochodzi się do sprzeczności. Rozwiązanie takie musi zawierać informację o tym, że sieć ma pięć węzłów ($W = 5$). Ponieważ każda para węzłów musi być połączona krawędzią, a mamy dziesięć takich par, to $K = 10$. Z twierdzenia Eulera wynika, że $S = K - W + 2 = 7$. Obszary sieci dualnej wytyczają pętle utworzone przez krawędzie, przy czym każdą parę węzłów łączy tylko jedna krawędź, a zatem takie pętle muszą się składać przynajmniej z 3 krawędzi. Ponieważ mamy 7 obszarów, to sieć musi zawierać przynajmniej 21 krawędzi… tyle tylko, że każdą krawędź policzyliśmy tu dwukrotnie, ponieważ rozdziela 2 obszary. Zatem sieć musi mieć przynajmniej $10\frac{1}{2}$ krawędzi. Liczba krawędzi musi być liczbą całkowitą, dochodzimy więc do wniosku, że sieć zawiera przynajmniej 11 krawędzi. Wcześniej jednak powiedzieliśmy, że nasza sieć ma tylko 10 krawędzi. Tym sposobem doszliśmy do sprzeczności logicznej, co dowodzi, że taka sieć nie może istnieć. Indyjski władca nie może podzielić swojego kraju w ten sposób.

Przeprowadzone właśnie rozumowanie napawa otuchą, ponieważ dzięki wykorzystaniu eleganckich metod topologicznych udało się nam udowodnić pewien interesujący fakt na temat map. Często spotyka się jednak błędne przekonanie, które podzielał chyba również De Morgan, że niemożność rozwiązania zagadki o pięciu indyjskich książętach dowodzi poprawności twierdzenia o czterech barwach. To nieprawda. Dowód może być błędny nawet wtedy, gdy płynie z niego wniosek prawidłowy lub taki, co do którego nie mamy pewności, że jest błędny. Jeśli w jakimś rzekomym dowodzie natrafiam na trójkąt o czterech bokach, to wiem, że mogę go już dalej nie czytać, ponieważ jest błędny. I nie ma znaczenia, co się

w nim dalej dzieje, ani jaki jest ostateczny wniosek. Nasze rozwiązanie problemu indyjskich książąt pokazuje tylko, że nie można obalić twierdzenia o czterech barwach w ten jeden, konkretny sposób. Nie wynika z tego, że nie może istnieć żaden *inny* sposób udowodnienia, iż twierdzenie to jest nieprawdziwe. Nie można wykluczyć istnienia innych przeszkód utrudniających pokolorowanie mapy za pomocą czterech barw. Możliwość istnienia pięciu obszarów stykających się ze sobą nawzajem jest tylko jedną z takich możliwości. Trudno na przykład z miejsca odrzucić możliwość, że istnieje jakaś bardzo skomplikowana mapa obejmująca 703 regiony o takim ich układzie, że jakkolwiek byśmy pokolorowali 702 obszary za pomocą czterech barw, to zawsze ostatni obszar będzie wymagał użycia piątego koloru. Obszar taki musiałby graniczyć przynajmniej z czterema innymi, ale to całkowicie możliwe, ponieważ układ tych obszarów nie musi spełniać wymagań podanych w zagadce o indyjskich książętach. Gdyby taka mapa istniała, stanowiłaby dowód na to, że cztery barwy nie wystarczają. Każdy dowód poprawności twierdzenia o czterech barwach musi wykluczyć tego typu przypadki. Stwierdzenie to pozostaje prawdziwe, nawet jeśli nie pokażę tu – bo nie potrafię – przykładu tego typu przeszkody.

Przez pewien czas wydawało się, że problem czterech barw popadł w zapomnienie, ale pojawił się ponownie w 1878 roku, gdy Arthur Cayley przypomniał o nim na spotkaniu Londyńskiego Towarzystwa Matematycznego. Choć nie wynika to z nazwy towarzystwa, którego założycielem był sam De Morgan, to jednak jest ono organizacją reprezentującą wszystkich brytyjskich (a przynajmniej angielskich) matematyków. Cayley spytał zebranych, czy komuś udało się znaleźć rozwiązanie. Wkrótce potem powtórzył swoje pytanie na łamach czasopisma naukowego „Nature". Rok później napisał dłuższy artykuł dla czasopisma „Proceedings of the Royal Geographical Society" wydawanego przez Królewskie Towarzystwo Geograficzne[31]. Zapewne wydawało się, że jest to odpowiednie miejsce na zadanie tego typu pytania, bo przecież problem dotyczy map. Niewykluczone nawet, że go o to poproszono. Nie był to jednak dobry wybór,

ponieważ kartografowie nie mieli powodów, by interesować się rozwiązaniem tej kwestii, no, może z wyjątkiem zwykłej ciekawości. Niestety, taki wybór czasopisma sprawił, że niewielu matematyków usłyszało o ukazaniu się artykułu. To wielka szkoda, ponieważ Cayley wyjaśniał w nim, dlaczego problem ten może być trudny do rozwiązania.

W rozdziale 1 powiedzieliśmy, że dowód przypomina nieco bitwę, a w wojsku odróżnia się taktykę od strategii. Taktyka dotyczy wygrywania lokalnych potyczek, natomiast strategia opisuje szersze aspekty całej kampanii. Taktyka zajmuje się szczegółowym ruchem poszczególnych oddziałów, strategia zaś polega na kreśleniu szerszych planów, w których na każdym etapie znajdzie się miejsce na wiele różnych decyzji taktycznych. W artykule Cayleya nie znajdziemy zbyt wiele taktyki, ale za to pojawia się tam bardzo niejasna wskazówka strategiczna, która w ostatecznym rozrachunku okazała się kluczowa dla uporania się z problemem czterech barw. Cayley zwrócił mianowicie uwagę na fakt, że dodawanie kolejnych obszarów nic nie daje, gdy stosuje się podejście standardowe, ale niewykluczone, że mogłoby przynieść efekty, gdyby zastosować jakieś mniej oczywiste rozumowanie.

Wyobraźmy sobie pewną mapę, z której usunięto jeden obszar – na przykład łącząc go z obszarem sąsiednim lub zmniejszając jego powierzchnię do pojedynczego punktu. Przypuśćmy, że tak zmienioną mapę można pokolorować czterema barwami. A teraz przywróćmy usunięty przed chwilą obszar. Jeśli dopisze nam szczęście, okaże się, że sąsiadujące z nim obszary używają tylko trzech barw. W takim wypadku wystarczy, że zamalujemy ten obszar czwartym kolorem. Cayley zauważył, że taka procedura nie zawsze musi zadziałać, ponieważ sąsiedzi przywróconego obszaru mogą już mieć cztery różne kolory. Nie oznacza to jednak, że znaleźliśmy się w sytuacji bez wyjścia. Można sobie z tym poradzić na dwa sposoby: może wybraliśmy nieodpowiedni obszar i wystarczy przeprowadzić całą operację ponownie dla innego obszaru lub możemy próbować inaczej pokolorować mapę przed przywróceniem usuniętego wcześniej obszaru.

Przyjmijmy na razie nieuzasadnione niczym założenie (to bardzo skuteczny sposób na wymyślanie nowych koncepcji, ale w którymś momencie założenia takie trzeba w końcu uzasadnić), że tego typu problem daje się zawsze rozwiązać. Z takiego założenia wynika, że daną mapę zawsze można pokolorować czterema barwami, pod warunkiem iż daje się w ten sposób pokolorować jakąś mniejszą mapę. Może się wydawać, że takie stwierdzenie nic nam nie daje, bo skąd wiadomo, że mniejszą mapę da się pokolorować czterema barwami? Wystarczy jednak zauważyć, że to samo stwierdzenie odnosi się również do tej mniejszej mapy, dzięki czemu dochodzimy do jeszcze mniejszej mapy… i tak dalej. W końcu uzyskamy mapę tak małą, że będą na niej tylko cztery obszary, a wtedy będziemy mieli pewność, że da się je pokolorować czterema barwami. Teraz wystarczy odwrócić cały proces, kolorując w każdym kroku nieco większą mapę, i na końcu dotrzemy do naszej mapy wyjściowej.

Takie rozumowanie to tak zwany dowód przez indukcję matematyczną. Jest to jedna ze standardowych metod dowodzenia, zdefiniowana z wykorzystaniem fachowej terminologii, dzięki czemu cały wywód logiczny można przeprowadzić w bardziej ścisły sposób. Zaproponowana przez Cayleya strategia dowodu stanie się bardziej zrozumiała, jeśli opiszemy całą metodę z wykorzystaniem logicznie równoważnego pojęcia „minimalnego kontrprzykładu". W tym kontekście „kontrprzykładem" jest każda hipotetyczna mapa, której nie można pokolorować czterema barwami. Powiemy, że jest ona minimalna, jeśli każda mapa o mniejszej liczbie obszarów *daje się* pokolorować czterema barwami. Jeśli istnieje choć jeden kontrprzykład, to musi również istnieć kontrprzykład minimalny – wystarczy wybrać kontrprzykład o najmniejszej możliwej liczbie obszarów. Zatem jeśli nie istnieje kontrprzykład minimalny, to w ogóle nie ma żadnego kontrprzykładu. A jeśli tak, to twierdzenie o czterech barwach musi być prawdziwe.

Proces indukcji matematycznej przebiega w tym wypadku następująco: przypuśćmy, że potrafimy udowodnić, iż pokolorowanie czterema barwami minimalnego kontrprzykładu jest możliwe, pod warunkiem że daje się tak pokolorować jakaś związana z nim mniejsza mapa. Wówczas ów minimalny kontrprzykład nie może być

w istocie kontrprzykładem. Ponieważ jest on minimalną mapą tego rodzaju, zatem *wszystkie* mniejsze mapy dają się pokolorować czterema barwami, a więc, na mocy naszego założenia, musi być możliwe pokolorowanie w ten sposób również naszej mapy wyjściowej. Wynika z tego, że nie istnieje żaden minimalny kontrprzykład, a więc w ogóle nie ma kontrprzykładów. Dzięki zastosowaniu takiego podejścia nie musimy już rozważać problemu wszystkich map. Możemy skupić uwagę tylko na hipotetycznych minimalnych kontrprzykładach i określeniu procedury redukcji – systematycznej procedury przekształcającej jakąś mniejszą mapę pokolorowaną czterema barwami w związaną z nią większą mapę, pokolorowaną w taki sam sposób.

Skąd całe to zamieszanie wokół minimalnych kontrprzykładów? Czy nie lepiej byłoby zająć się po prostu kontrprzykładami? To kwestia podejścia. Nawet jeśli początkowo nie wiemy, czy kontrprzykłady w ogóle istnieją, to jedną z paradoksalnych, ale użytecznych cech tej strategii jest to, że możemy dość dużo powiedzieć na temat tego, jak powinny wyglądać kontrprzykłady minimalne, jeśli istnieją.

Wymaga to biegłości w prowadzeniu wywodu logicznego operującego hipotezami – to podstawowa umiejętność, jaką musi opanować każdy matematyk. Abyśmy mogli się przekonać, jak przebiega taki proces, przeprowadźmy dowód twierdzenia o *sześciu* barwach. W tym celu musimy zastosować tę samą sztuczkę, której użyliśmy przy omawianiu zagadki o pięciu książętach, i sformułować wszystko za pomocą sieci dualnej. W takim ujęciu obszary stają się punktami, a problem czterech barw przekształca się w inne pytanie: Czy można pokolorować czterema barwami *węzły* sieci o nieprzecinających się krawędziach w taki sposób, żeby dwa węzły połączone krawędzią miały zawsze różny kolor? W taki sam sposób można sformułować pytanie na temat dowolnej innej liczby kolorów.

Aby pokazać skuteczność podejścia z minimalnym kontrprzykładem, posłużmy się nim do udowodnienia, że każdą sieć na płaszczyźnie można pokolorować sześcioma barwami. Tak jak poprzednio, naszym najważniejszym narzędziem będzie wzór Eulera.

Zdefiniujmy najpierw pojęcie sąsiedztwa: sąsiadem danego węzła sieci dualnej jest każdy węzeł połączony z nim krawędziami. Węzeł może mieć wielu sąsiadów lub zaledwie kilku. Można jednak pokazać, że ze wzoru Eulera wynika, iż niektóre węzły muszą mieć niewielu sąsiadów. Mówiąc dokładniej, na płaszczyźnie nie można narysować sieci, w której wszystkie węzły miałyby sześciu lub więcej sąsiadów. Dowód tego twierdzenia zamieściłem w uwagach, żeby nie burzył nam toku rozumowania[32]. Fakt ten jest odpowiednim punktem zaczepienia, który umożliwi nam rozłożenie problemu na prostsze elementy. Wróćmy więc do twierdzenia o sześciu barwach i załóżmy, że istnieje hipotetyczny minimalny kontrprzykład. Jest on siecią, której nie można pokolorować sześcioma barwami, ale każda sieć mniejsza daje się już w ten sposób pokolorować. Udowodnimy teraz, że taka sieć nie może istnieć. Uwzględniając wspomniany przed chwilą wniosek płynący ze wzoru Eulera, możemy zauważyć, że nasza sieć musi zawierać przynajmniej jeden węzeł posiadający pięciu lub mniej sąsiadów. Usuńmy go na chwilę, a razem z nim wszystkie krawędzie łączące go z sąsiadami. W ten sposób otrzymujemy sieć o mniejszej liczbie węzłów. Teraz zauważmy, że z faktu, iż wyjściowa sieć była minimalnym kontrprzykładem, wynika, że taką pomniejszoną sieć można pokolorować sześcioma barwami. (Właśnie w tym miejscu utknęlibyśmy, gdyby nasza wyjściowa sieć nie była *minimalnym* kontrprzykładem). Przywróćmy teraz usunięty węzeł i jego krawędzie. Węzeł ten ma najwyżej pięciu sąsiadów, pozostaje więc nam szósty, niewykorzystany kolor – pokolorujmy nim usunięty węzeł. W ten sposób pokolorowaliśmy sześcioma barwami naszą sieć wyjściową – z tego jednak wynika, że nie jest ona kontrprzykładem. Zatem dla twierdzenia o sześciu barwach nie mogą istnieć minimalne kontrprzykłady, a to oznacza, że twierdzenie jest prawdziwe.

To pokrzepiające. Do tej chwili nie potrafiliśmy odrzucić możliwości, że jakieś mapy mogą wymagać użycia 20, 703 lub milionów kolorów. Teraz wiemy, że tego typu mapy są równie prawdziwe jak kwiat paproci. Pokazaliśmy, że określona, ograniczona liczba kolorów wystarcza do pokolorowania *każdej* mapy. Ten prawdziwy tryumf

pojęcia minimalnego kontrprzykładu zachęcił matematyków do podjęcia prób uściślenia wywodu w nadziei zastąpienia sześciu kolorów pięcioma, a może nawet, kto wie, czterema.

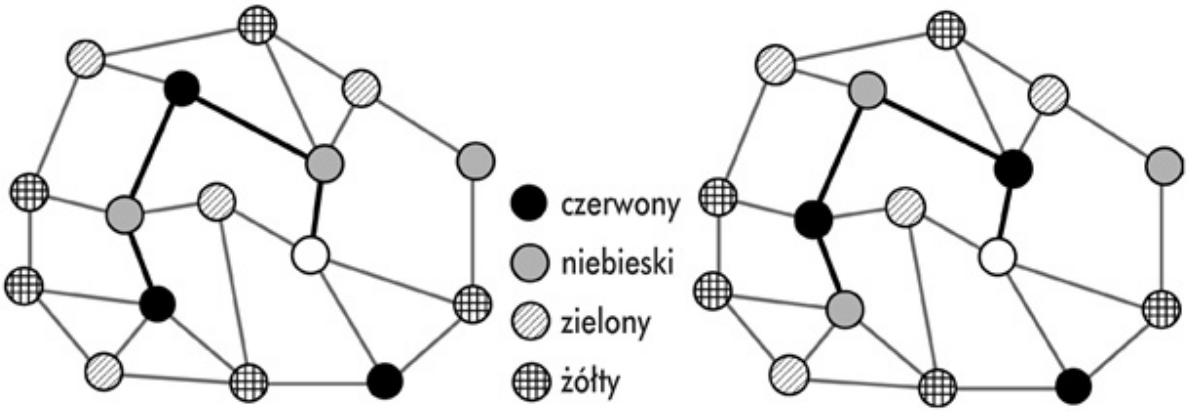
Jeśli chodzi o przedstawianie długich wywodów, to nikt nie może w tej materii rywalizować z prawnikami. Na spotkaniu, na którym Cayley wspomniał o problemie czterech barw, był również pewien adwokat, Alfred Kempe. Kempe studiował kiedyś matematykę u Cayleya na uniwersytecie w Cambridge i wciąż żywo interesował się tą dziedziną. Po roku Kempe był przekonany, że udało mu się rozwiązać ten problem, i w 1879 roku opublikował swoje rozwiązanie w świeżo założonym czasopiśmie „American Journal of Mathematics". Rok później przedstawił uproszczony dowód, w którym poprawił kilka wcześniejszych błędów. W jego artykule można znaleźć następującą uwagę:

> Bardzo mała zmiana w jednej części mapy może wywołać konieczność ponownego pokolorowania jej całej. Po dość żmudnych poszukiwaniach udało mi się […] odkryć pewien słaby punkt, który okazał się prosty do zaatakowania.

Opiszmy idee Kempe'a za pomocą sieci dualnej. Również on wyszedł od wzoru Eulera i wynikającego z niego wniosku o możliwości istnienia węzła z trzema, czterema lub pięcioma sąsiadami. (Węzeł z dwoma sąsiadami leży po prostu w środku linii i nie wnosi nic do sieci lub mapy – takie węzły można bezpiecznie pominąć).

Jeśli istnieje jakiś węzeł z trzema sąsiadami, to możemy zastosować taką samą procedurę jak w naszym wcześniejszym dowodzie twierdzenia o sześciu barwach. Usuńmy ten węzeł i związane z nim krawędzie, pokolorujmy trzema barwami to, co zostało, a następnie przywróćmy usunięty węzeł z krawędziami i pokolorujmy go czwartym, niewykorzystanym kolorem. Możemy więc założyć, że żaden z węzłów nie ma trzech sąsiadów.

Jeśli istnieje jakiś węzeł z czterema sąsiadami, to taka taktyka zawodzi, ponieważ po pierwszym kolorowaniu może już nam nie zostać żaden niewykorzystany kolor. Kempe wymyślił sprytny sposób na ominięcie tej przeszkody: usuńmy mimo wszystko ten węzeł, ale zaraz potem zmieńmy kolory w uzyskanej w ten sposób mniejszej sieci w taki sposób, żeby dwa z owych czterech sąsiednich węzłów miały taki sam kolor. Po takiej zmianie pokolorowanie sąsiadów usuniętego węzła wymaga użycia najwyżej trzech barw, pozostaje więc jeden niewykorzystany kolor dla węzła, który usunęliśmy. Pomysł Kempe'a na ponowne pokolorowanie sieci opiera się na spostrzeżeniu, że dwa z tych sąsiednich węzłów muszą być różnego koloru – powiedzmy, czerwonego i niebieskiego – natomiast pozostałe dwa mogą być zielone lub żółte. Jeśli oba są żółte lub zielone, to automatycznie możemy użyć pozostałej barwy do pokolorowania usuniętego węzła. Możemy więc założyć, że jeden z tych węzłów jest zielony, a drugi – żółty. Znajdźmy teraz wszystkie węzły, które można połączyć z węzłem niebieskim ciągiem krawędzi, przechodzącym tylko przez węzły niebieskie i czerwone. Nazwijmy ten ciąg niebiesko-czerwonym łańcuchem Kempe'a[33]. Z definicji wynika, że każdy sąsiad węzłów w łańcuchu Kempe'a, który sam do łańcucha nie należy, jest albo zielony, albo żółty, ponieważ sąsiad o kolorze niebieskim lub czerwonym trafiłby do łańcucha. Zwróćmy teraz uwagę, że zamiana kolorów węzłów w takim łańcuchu – czerwone na niebieskie, a niebieskie na czerwone – prowadzi do powstania sieci pokolorowanej inaczej, ale wciąż spełniającej podstawowe założenie, że sąsiednie węzły muszą mieć różne kolory (zob. ryc. 11).



Ryc. 11. Zamiana kolorów w łańcuchu Kempe'a (pogrubione linie) związanym z węzłem czwartego stopnia (zaznaczonym na rysunku białym kolorem), którego sąsiedzi mają cztery różne kolory. Po lewej: Wyjściowy układ kolorów. Po prawej: Po zamianie kolorów biały węzeł można pokolorować na niebiesko.

Jeśli czerwony sąsiad naszego usuniętego węzła nie należy do tego łańcucha, dokonajmy takiej zamiany kolorów. W jej wyniku niebieski sąsiad usuniętego węzła stanie się czerwony, a czerwony nie zmieni koloru. W efekcie sąsiedzi usuniętego węzła mają teraz najwyżej trzy różne kolory: czerwony, zielony i żółty. W ten sposób pozostaje nam kolor niebieski, którym możemy pomalować usunięty węzeł, i zadanie wykonane. Może się jednak zdarzyć, że niebiesko-czerwony łańcuch zakręca i dochodzi do niebieskiego sąsiada usuniętego węzła. W takiej sytuacji zostawiamy niebiesko-czerwony łańcuch w spokoju i stosujemy taką samą sztuczkę do sąsiadów o kolorze zielonym i żółtym. Zacznijmy od węzła zielonego i utwórzmy żółto-zielony łańcuch Kempe'a. Ten łańcuch *nie może* dochodzić do żółtego sąsiada, ponieważ nie pozwala na to łańcuch niebiesko-czerwony. Teraz wystarczy zamienić w łańcuchu kolor żółty na zielony, a zielony na żółty, i zadanie wykonane.

Pozostaje nam więc już tylko ostatni przypadek do rozważenia, czyli sieć, w której nie ma węzłów z trzema lub czterema sąsiadami, ale przynajmniej jeden węzeł ma pięciu sąsiadów. Kempe zaproponował podobną, ale bardziej zawiłą regułę ponownego

kolorowania i wydawało się, że tym sposobem udało mu się rozwiązać zadanie również w tym przypadku. Wniosek: twierdzenie o czterech barwach jest prawdziwe, a dowiódł tego Kempe. Informacja o tym trafiła nawet do prasy – w amerykańskim czasopiśmie „The Nation" wspomniano o jego osiągnięciu w artykule przeglądowym.

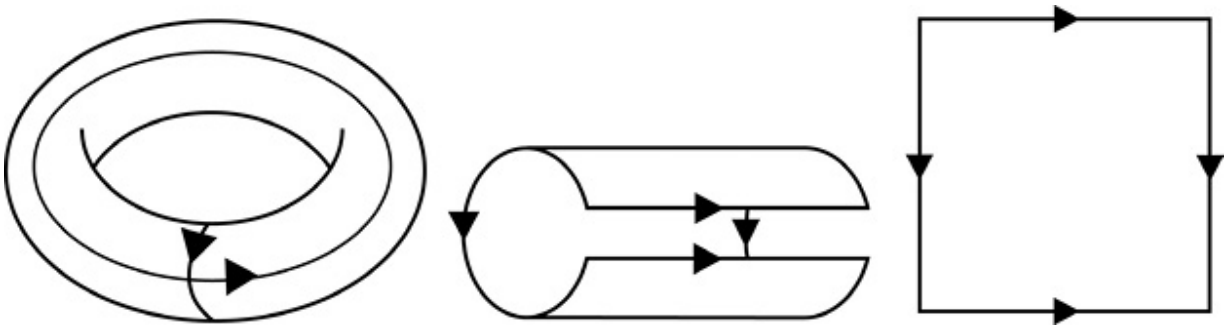
Wydawało się, że dowód Kempe'a rozwiązał problem. Dla większości matematyków sprawa była zamknięta. Jedynie Peter Guthrie Tait nadal publikował artykuły poświęcone temu zagadnieniu, szukając prostszego dowodu – tym sposobem dokonał kilku użytecznych odkryć, ale nie udało mu się uprościć dowodu.

W tym momencie w naszej opowieści pojawia się Percy Heawood, wykładowca matematyki na uniwersytecie w Durham, znany wszystkim jako „Kocur" ze względu na sumiaste wąsy. O problemie czterech barw dowiedział się już na studiach w Oksfordzie od Henry'ego Smitha, tamtejszego profesora geometrii. Smith powiedział mu, że twierdzenie to być może jest prawdziwe, ale nie zostało jeszcze udowodnione, i Heawood postanowił spróbować własnych sił. Gdy zabrał się do roboty, wpadł mu w ręce artykuł Kempe'a, próbował więc zrozumieć przedstawiony w nim dowód. Wynik swoich wysiłków opublikował w 1889 roku w artykule zatytułowanym *Twierdzenie o kolorowaniu map*. Wyraził w nim żal z powodu tego, że jego artykuł jest bardziej „destruktywny niż konstruktywny, ponieważ zostanie w nim wykazane, że w tym dowodzie, najwyraźniej powszechnie uznawanym za poprawny, tkwi usterka". Kempe popełnił błąd.

Był to trudny do wykrycia błąd w metodzie ponownego kolorowania sieci w wypadku, gdy usunięty węzeł miał pięciu sąsiadów. Opisana przez Kempe'a procedura mogła czasami prowadzić do ponownej zmiany koloru danego węzła na skutek efektu domina wywołanego jakąś późniejszą zmianą. Kempe uznał jednak, że raz zmieniony kolor nie ulega już później zmianie. Heawood znalazł sieć, dla której procedura Kempe'a działa niepoprawnie, okazało się więc, że dowód jest błędny. Kempe szybko

przyznał się do pomyłki i dodał, że nie udało mu się „naprawić tego defektu”. Twierdzenie o czterech barwach było znów do wzięcia.

Heawood ustalił, że z tej katastrofy można jednak coś ocalić: Kempe mógł się pocieszyć tym, że jego metoda pozwala poprawnie dowieść prawdziwości twierdzenia o pięciu barwach. Heawood zainteresował się również dwoma uogólnieniami tego problemu: przykładem imperiów, których obszary składają się z kilku rozłącznych obszarów i wszystkie je trzeba oznaczyć tym samym kolorem, oraz mapami na bardziej skomplikowanych powierzchniach. Analogicznie sformułowane pytanie dla powierzchni sferycznej ma taką samą odpowiedź jak w przypadku mapy na płaszczyźnie. Wyobraźmy sobie mapę na powierzchni kuli, którą obracamy tak, żeby biegun północny znalazł się we wnętrzu jednego z obszarów. Jeśli teraz usuniemy ten biegun, to możemy rozwinąć przekłutą sferę i uzyskamy przestrzeń topologicznie równoważną nieskończonej płaszczyźnie. Obszar zawierający biegun północny przekształci się w wówczas w nieskończenie duży obszar otaczający pozostałą część mapy. Istnieją jednak inne, bardziej interesujące powierzchnie. Należy do nich torus – bryła przypominająca kształtem obwarzanek (zob. ryc. 12, po lewej stronie).



Ryc. 12. Po przecięciu i rozwinięciu torusa uzyskujemy kwadrat.

Jest pewien dobry sposób na wyobrażenie sobie torusa, który pozwala wszystko uprościć. Jeśli rozetniemy torus wzdłuż dwóch zamkniętych krzywych (zob. ryc. 12, w środku), to po jego rozwinięciu otrzymamy kwadrat (ryc. 12, po prawej). Takie

przekształcenie zmienia topologię torusa, ale możemy to obejść, „utożsamiając” ze sobą przeciwległe boki kwadratu. W zasadzie (można podać ścisłą definicję opisującą tę ideę dokładnie) umawiamy się, że będziemy traktować odpowiednie punkty tych boków tak, jak gdyby były jednym i tym samym punktem. Aby zrozumieć, jak to działa, spójrzmy na te rysunki w odwrotnej kolejności. Po zwinięciu kwadratu przeciwległe boki faktycznie zostają ze sobą sklejone. Ów sprytny pomysł polega na tym, że tak naprawdę nie musimy wcale zwijać kwadratu, żeby połączyć ze sobą jego boki. Możemy przeprowadzać wszystkie operacje na kwadracie, pamiętając cały czas, że przeciwległe boki są tożsame. Wszystkie działania na torusie, na przykład narysowanie na jego powierzchni jakiejś krzywej, można przedstawić za pomocą odpowiadających im działań na kwadracie.

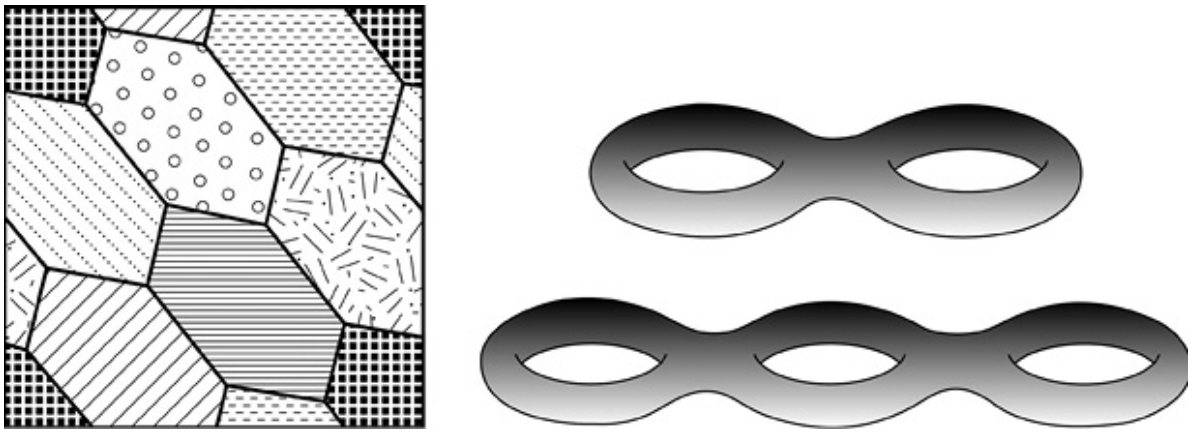
Heawood udowodnił, że do pokolorowania mapy na torusie potrzeba i wystarcza siedem kolorów. Na rycinie 13, po lewej stronie, pokazano na kwadracie przedstawiającym torus w opisany przed chwilą sposób, że do pokolorowania mapy potrzeba co najmniej siedmiu kolorów. Zwróćmy uwagę, jak pasują do siebie obszary leżące na przeciwległych krawędziach. Istnieją również powierzchnie przypominające torus, mające jednak więcej otworów (zob. ryc. 13, po prawej). Liczbę otworów nazywamy genusem powierzchni i oznaczamy literą *g*. Heawood wysunął hipotezę, w której przedstawił wzór na liczbę kolorów potrzebnych na torusie z *g* otworami, dla $g \geq 1$: jest to najmniejsza liczba całkowita mniejsza lub równa wartości

$$\frac{7+\sqrt{48g+1}}{2}$$

Dla kolejnych wartości *g* z przedziału od 1 do 10 wzór ten daje wartości:

7 8 9 10 11 12 12 13 13 14.

Liczba kolorów wynikająca z tego wzoru rośnie wolniej niż genus i często dodanie kolejnego otworu w torusie nie robi różnicy. To zaskakujące, ponieważ każdy dodatkowy otwór zwiększa możliwości wymyślenia bardziej skomplikowanych map.



Ryc. 13. Po lewej: Siedmiokolorowa mapa na torusie. Torus jest tu przedstawiony za pomocą kwadratu i należy sobie wyobrazić, że jego przeciwległe boki są „zwinięte" i sklejone ze sobą. Poszczególne obszary mapy na przeciwległych bokach muszą do siebie pasować. Po prawej: torusy z dwoma i trzema otworami.

Wzór Heawooda nie wziął się z powietrza. Powstał w wyniku uogólnienia metody, którą posłużyliśmy się do udowodnienia twierdzenia o sześciu barwach na płaszczyźnie. Heawood dowiódł, że taka liczba kolorów będzie zawsze wystarczająca. Przez wiele lat otwartym pytaniem było, czy liczba ta mogłaby być mniejsza. Matematycy znaleźli przykłady dla mniejszych wartości genusu, które sugerowały, że oszacowanie Heawooda może dawać najlepszą możliwą wartość. W 1968 roku po długich badaniach Gerhard Ringel i John W.T. (Ted) Youngs uzupełnili brakujące szczegóły dowodu pokazującego, że tak właśnie jest. W swojej pracy wykorzystali własne wcześniejsze dokonania, a także odkrycia kilku innych matematyków. Zastosowali metody kombinatoryczne, wykorzystujące specjalne rodzaje sieci, które są tak skomplikowane, że ich omówienie wymagałoby napisania oddzielnej książki[34].

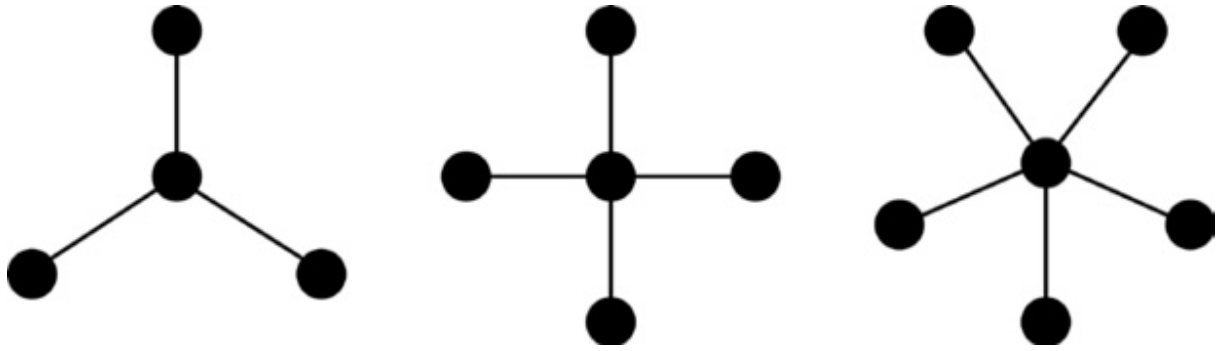
Gdy $g = 0$, czyli w przypadku map na sferze, ze wzoru Heawooda wynika, że potrzebne są cztery kolory, ale przedstawiony przez Heawooda dowód, że taka liczba barw wystarcza, nie działa dla sfery. Mimo imponujących postępów w zakresie badań powierzchni z co najmniej jednym otworem, oryginalne sformułowanie problemu czterech barw wciąż nie miało rozwiązania. Nieliczni matematycy, którzy byli gotowi poświęcić czas na badanie tej kwestii, przygotowali się do długiego oblężenia. Problem czterech barw przypominał dobrze bronioną twierdzę. Uczeni mieli nadzieję, że będą mogli konstruować coraz lepsze machiny oblężnicze i powoli kruszyć mury zamku tak długo, aż rozsypią się w pył. Budowali więc kolejne machiny, ale mury pozostały niewzruszone. Atakującym udało się jednak zgromadzić wiele informacji na temat tego, jak nie należy rozwiązywać tego problemu i jakie przeszkody wydają się nie do uniknięcia. Z porażek zaczęła się powoli wyłaniać ambitna strategia. Było to naturalne rozszerzenie metod Kempe'a i Heawooda, opierające się na trzech krokach. Opiszmy to podejście z wykorzystaniem pojęcia sieci dualnej – obecnie właśnie tak się je najczęściej przedstawia:

1. Załóżmy, że istnieje minimalny kontrprzykład.
2. Znajdźmy listę wszystkich konfiguracji niemożliwych do uniknięcia, czyli listę mniejszych sieci charakteryzujących się tym, że każdy minimalny kontrprzykład musi zawierać jakiś znajdujący się na niej element.
3. Udowodnijmy, że wszystkie nieuniknione konfiguracje są redukowalne. Innymi słowy, pokażmy, że jeśli można pokolorować czterema barwami mniejszą sieć uzyskaną w wyniku usunięcia węzła z konfiguracji nie do uniknięcia, to kolory te można zawsze pozmieniać w taki sposób, że po przywróceniu usuniętego węzła również cała sieć wyjściowa będzie pokolorowana czterema barwami.

Zrealizowanie tych trzech kroków będzie równoznaczne z udowodnieniem, że minimalny kontrprzykład nie istnieje. Gdyby

bowiem istniał, musiałby zawierać jedną z nieuniknionych konfiguracji. Pozostała część sieci jest jednak mniejsza, zatem z faktu, że mamy do czynienia z *minimalnym* kontrprzykładem, wynika, iż taką mniejszą sieć można pokolorować czterema barwami. Teraz wystarczy skorzystać z faktu redukowalności, by dowieść, że wobec tego również całą sieć wyjściową można pokolorować czterema barwami. W ten sposób doszliśmy do sprzeczności.

Idąc tym tropem, Kempe znalazł poprawną listę konfiguracji niemożliwych do uniknięcia – są to węzły z wychodzącymi z nich trzema, czterema i pięcioma krawędziami (zob. ryc. 14). Udało mu się również dowieść, że pierwsze dwie z tych konfiguracji są redukowalne. Pomylił się tylko, dowodząc, że także trzecia konfiguracja jest redukowalna. To nieprawda. Trzeba zastosować inne rozwiązanie: należy zastąpić tę złą konfigurację jakąś dłuższą listą, upewniając się jednak, że jest to zbiór nieunikniony. Trzeba to zrobić tak, by każda z konfiguracji z tej nowej listy była redukowalna. Innymi słowy, należy poszukać nieuniknionej listy redukowalnych konfiguracji. Jeśli to się nam uda, udowodnimy twierdzenie o czterech barwach.



Ryc. 14. Lista nieuniknionych konfiguracji sporządzona przez Kempe'a

Nie było gwarancji, że taka lista istnieje, ale należało podjąć próbę realizacji tej strategii, tym bardziej że nikt nie miał lepszego pomysłu. Taka lista musiałaby się jednak charakteryzować dziwnym wewnętrznym napięciem. Z jednej strony obszerniejsza lista daje

większe szanse na to, że będzie nieunikniona, a to bardzo dobrze. Z drugiej jednak strony dłuższa lista oznacza mniejsze prawdopodobieństwo tego, że każda znajdująca się na niej konfiguracja będzie redukowalna. Nawet jeśli tylko jeden element z listy okaże się nieredukowalny, to cały dowód legnie w gruzach, a ryzyko, że tak się stanie, rośnie wraz z długością listy. To bardzo niedobrze. Z *trzeciej* jednak strony… dłuższa lista daje więcej okazji do wybrania redukowalnych konfiguracji, a to dobrze. Z czwartej strony, przy dłuższej liście trzeba włożyć więcej pracy w dokończenie dowodu, a to źle. I wreszcie, z piątej strony, matematycy i tak nie mieli żadnych dobrych metod realizacji tych zamierzeń, a to po prostu fatalnie.

Właśnie takie przeszkody sprawiają, że dany problem zasługuje na miano wielkiego.

Przez pewien okres udawało się od czasu do czasu odłupać jakiś fragment muru otaczającego zamek, ale taki uszczerbek ani trochę nie osłabiał solidności konstrukcji całej twierdzy. Tymczasem obserwujący to matematycy z głównego nurtu badań ziewali z nudów, jeśli w ogóle zwracali uwagę na te starania. W końcu jednak znalazł się ktoś, komu udało się skonstruować lepszy taran – tym sprytnym uczonym był Heinrich Heesch. Jego ważny wkład w te prace polegał na opracowaniu systematycznej metody dowodzenia, że dana konfiguracja jest redukowalna. Swoje podejście nazwał „rozładowaniem". Jego koncepcję można z grubsza porównać do sytuacji, w której węzły sieci są obdarzone ładunkiem elektrycznym, a krawędzie są przewodnikami umożliwiającymi swobodny przepływ prądu.

Jednak nawet przy użyciu tej metody znalezienie nieuniknionego zbioru redukowalnych konfiguracji wymagałoby przeprowadzenia niezwykle żmudnych obliczeń. Pojedyncze konfiguracje powinny być dość małe, ale należało się spodziewać, że będzie ich naprawdę dużo. Heesch nie ustawał w wysiłkach i w 1948 roku wygłosił cykl wykładów, z których wynikało, że należy przeanalizować około 10 000 konfiguracji. W tym okresie udało mu się już dowieść, że 500 proponowanych konfiguracji daje się zredukować. Wykładom przysłuchiwał się młody student Wolfgang Haken, który później

wyznał, że niewiele tak naprawdę wówczas z nich zrozumiał, ale niektóre z najważniejszych kwestii wbiły mu się w pamięć. Haken zajął się później topologią i udało mu się dokonać ważnych odkryć z dziedziny teorii węzłów. Dzięki temu nabrał śmiałości, żeby zmierzyć się z hipotezą Poincarégo (powiemy o tym w rozdziale 10). Podejście, które zastosował tym razem, wymagało sklasyfikowania i przeanalizowania 200 przypadków – z czego 198 udało mu się rozwiązać, ale z pozostałymi dwoma zmagał się przez trzynaście lat. W końcu się poddał i zainteresował problemem czterech barw. Nie ulega wątpliwości, że Haken lubił stawiać czoło trudnym wyzwaniom, ale obawiał się, że podobna sytuacja jak przy pracy nad hipotezą Poincarégo może go spotkać również podczas analizy 10 000 konfiguracji Heescha. Wyobraźmy sobie, jakby się czuł, gdyby udało mu się uporać z 9998 konfiguracjami, po czym utknąłby na ostatnich dwóch. Postanowił więc poprosić o radę Heescha i w 1967 roku zaprosił go do złożenia wizyty na Uniwersytecie Illinois, gdzie pracował.

W tamtym okresie po raz pierwszy zaczęto wykorzystywać komputery do prowadzenia poważnych badań matematycznych, ale były to ogromne urządzenia, zamknięte w specjalnych pomieszczeniach, a nie przyrządy mieszczące się na biurku czy w teczce. Haken zastanawiał się, czy mogłyby im w czymś pomóc. Heesch już wcześniej wpadł na ten pomysł i dokonał oszacowania złożoności analizowanego problemu. Z jego rozważań wynikało, że najlepszy komputer, do jakiego miał dostęp, nie byłby w stanie poradzić sobie z tym zadaniem. Na Uniwersytecie Illinois działał jednak dużo lepszy superkomputer ILLIAC IV, Haken więc poprosił o przyznanie mu czasu na tej maszynie. Niestety okazało się, że superkomputer jest obłożony, i poradzono mu, żeby spróbował uzyskać dostęp do komputera Cray 6600 działającego w Laboratorium Brookhaven na Long Island. Dyrektorem centrum komputerowego był tam wówczas Yoshio Shimamoto, którego już od dawna fascynował problem czterech barw – dzięki temu szczęśliwemu zbiegowi okoliczności Heesch i Haken uzyskali dostęp do komputera.

Komputer spełniał pokładane w nim nadzieje, ale Haken zaczął się zastanawiać, czy nie można by go jeszcze lepiej wykorzystać. Heesch i Haken znajdowali dużo redukowalnych konfiguracji i mieli nadzieję, że uda im się zestawić nieuniknioną listę, ale przy okazji tracili dużo czasu na analizę obiecująco wyglądających konfiguracji, które okazywały się jednak nieredukowalne. Może więc bardziej opłacałoby się postępować odwrotnie, to znaczy przyjąć, że głównym celem jest uzyskanie nieuniknionych konfiguracji, a redukowalność sprawdzać dopiero w dalszej kolejności? Oczywiście należało wybierać takie konfiguracje, które prawdopodobnie uda się zredukować, ale to podejście wydawało się lepsze. Wtedy jednak Cray w Laboratorium Brookhaven został już przeznaczony do realizacji ważniejszych zadań. Co gorsza, kilku specjalistów powiedziało Hakenowi, że metody, które chciał wykorzystać, zupełnie nie nadają się do przekształcenia w programy komputerowe. Uwierzył im i wygłosił wykład, podczas którego wyznał, że już dawno temu doszedł do wniosku, iż problemu czterech barw nie da się rozwiązać bez komputerów, a teraz wydaje się, że nawet komputery nie są w stanie pomóc w realizacji tego zadania. Postanowił się poddać.

Na sali znalazł się jednak doświadczony programista Kenneth Appel, który po wykładzie powiedział Hakenowi, że wspomniani przez niego specjaliści zapewne starali się go tylko zniechęcić, ponieważ przygotowanie takich programów wymagałoby dużo pracy, a efekt końcowy wcale nie był przesądzony. Sam Appel uważał, że nie ma takiego problemu matematycznego, którego nie dałoby się wyrazić za pomocą programu. Kluczową kwestią jest to, czy taki program da jakieś sensowne wyniki w rozsądnym czasie. Haken i Appel połączyli siły. Stosowana przez nich strategia ewoluowała, ponieważ modyfikacje metody rozładowywania wymagały wprowadzania zmian w programie, a usprawnienia programu pociągały za sobą konieczność zmian w metodzie. Tym sposobem doszli do nowego pojęcia konfiguracji „geograficznie odpowiednich”, czyli takich, które pozbawione są pewnych nieprzyjemnych cech uniemożliwiających ich redukcję. Szanse na to, że takie konfiguracje będą redukowalne, były dużo większe, a własność decydującą o tym,

że dana konfiguracja jest geograficznie odpowiednia, można było łatwo sprawdzić. Appel i Haken postanowili udowodnić teoretycznie, a nie z wykorzystaniem komputera, że musi istnieć nieunikniona lista konfiguracji geograficznie odpowiednich. Udało im się tego dokonać w 1974 roku.

Uzyskany wynik napawał otuchą, ale przeczuwali, co się zapewne zaraz stanie: niektóre z ich geograficznie odpowiednich konfiguracji okażą się nieredukowalne, a tym samym będą musieli je usunąć i zastąpić jeszcze dłuższą, bardziej skomplikowaną listą. Takie obliczenia będą goniły własny ogon i zakończą się sukcesem jedynie wtedy, gdy go złapią. Zamiast tracić lata na bezowocne wysiłki, postanowili oszacować z grubsza, ile czasu może zabrać tego typu proces. Uzyskane liczby napawały umiarkowanym optymizmem, zabrali się więc do roboty. Teoria i metody obliczeniowe wzajemnie się inspirowały i zmieniały. Czasami uczonym wydawało się, że komputer myśli samodzielnie i „odkrywa" użyteczne cechy konfiguracji. W pewnym momencie dział administracyjny uniwersytetu zafundował sobie nowy, bardzo wydajny komputer – lepszy od tych, którymi posługiwali się pracujący na uniwersytecie uczeni. Po kilku protestach i ostrej wymianie zdań badaczom udostępniono połowę czasu na nowym komputerze. Nieustannie zmieniająca się lista nieuniknionych konfiguracji Appela i Hakena w końcu się ustabilizowała na poziomie 2000 pozycji. W czerwcu 1976 roku komputer wypluł z siebie wynik ostatniej weryfikacji redukowalności i dowód był gotowy. Wieść o tym trafiła do prasy – jako pierwszy poinformował o osiągnięciu „The Times", a potem wiadomość szybko rozeszła się po całym świecie.

Appel i Haken musieli się jeszcze upewnić, że nie popełnili żadnej głupiej pomyłki, ale nie mieli na to zbyt dużo czasu, ponieważ kilka innych zespołów deptało im już po piętach. W lipcu byli już przekonani, że ich metoda działa, i ogłosili swój dowód oficjalnie, rozsyłając wśród matematyków wstępną wersję artykułu – zgodnie z obowiązującym wówczas zwyczajem autorzy kopiowali jak najtańszym kosztem pierwszą wersję artykułu i przesyłali ją do innych instytutów jeszcze przed ukazaniem się ich pracy w czasopiśmie. W tamtych czasach proces od przesłania artykułu

matematycznego do redakcji do jego faktycznego ukazania się w czasopiśmie zabierał od roku do dwóch lat. Aby więc uniknąć opóźniania dalszego postępu prac, uczeni znaleźli szybszy sposób na przekazywanie swoich wyników innym naukowcom. Obecnie wersje wstępne artykułów umieszcza się w Internecie. Artykuły na tym etapie zawsze mają charakter tymczasowy – zanim praca ukaże się oficjalnie, musi przejść zwycięsko proces recenzji. Wstępne wersje artykułu pomagają w tym, ponieważ każdy może je przeczytać, poszukać w nich błędów lub możliwości wprowadzenia jakichś poprawek – i poinformować o tym autorów. W istocie z tego właśnie powodu wersja ostateczna pracy nierzadko znacznie się różni od wersji wstępnej.

Ostateczna wersja dowodu wymagała tysiąca godzin obliczeń komputerowych i uwzględnienia 487 reguł rozładowywania. Wynik ukazał się w dwóch artykułach, do których dołączono 450-stronicowy dodatek zawierające wszystkie 1482 konfiguracje. W tamtym czasach był to prawdziwy majstersztyk.

Jednak społeczność matematyków przyjęła ten dowód z niejasnym uczuciem rozczarowania. Nie było to rozczarowanie wynikiem ani nadzwyczajną liczbą przeprowadzonych obliczeń. Rozczarowująca była metoda. W latach siedemdziesiątych dowody matematyczne były czymś, co można było wymyślić i zweryfikować ręcznie, na kartce papieru. Jak powiedzieliśmy w rozdziale 1, dowód jest opowieścią, której wątek przekonuje nas o prawdziwości danego stwierdzenia. Jednak w tym dowodzie nie było żadnego wątku. A jeśli był, to w samym jego środku zionęła wielka dziura:

> Za górami, za lasami, żyła sobie piękna hipoteza. Matka ostrzegała ją, żeby nigdy nie wchodziła do mrocznego, niebezpiecznego lasu. Pewnego jednak razu mała Hipoteza o Czterech Barwach wymknęła się chyłkiem z domu i weszła do nieuniknionego lasu. Wiedziała, że jeśli każda konfiguracja lasu okaże się redukowalna, to zdobędzie dowód i zmieni się w małe Twierdzenie o Czterech Barwach. Wtedy na pewno

trafi do różnych czasopism i być może usłyszy o niej nawet sam książę Fields. Błąkając się po lesie, natrafiła na komputer, cały z piernika, a w jego środku spotkała wilka przebranego za programistę. Usłyszawszy jej historię, wilk szybko ją zapewnił: „Tak, wszystkie te konfiguracje są redukowalne” i żyli potem razem długo i szczęśliwie.

Nie, to nieprawdopodobne. Trochę sobie tu zażartowałem, ale dziura w tej opowieści jest taka sama jak dziura w dowodzie Appela i Hakena, a przynajmniej większość matematyków uważała, że tak jest. *Jak się upewnić, że wilk ma rację?*

Trzeba uruchomić własny program komputerowy i sprawdzić, czy uzyskamy takie same wyniki. Jednak bez względu na to, ile razy powtórzymy taki proces, nie będzie on tak samo przekonujący jak, powiedzmy, nasz dowód tego, że wyszczerbionej szachownicy nie można pokryć kostkami domina. Takiego komputerowego dowodu nie można objąć w całości umysłem. Nie zdołamy sprawdzić ręcznie wszystkich obliczeń, nawet gdybyśmy żyli miliard lat. Jest jeszcze gorzej: nie uwierzylibyśmy w uzyskaną odpowiedź, nawet gdybyśmy mogli przeprowadzić takie obliczenia. Ludzie popełniają błędy, a w ciągu miliarda lat są zdolni do popełnienia naprawdę dużej ich liczby.

Komputery natomiast zwykle się nie mylą. Jeśli po przeprowadzeniu naprawdę skomplikowanych obliczeń arytmetycznych komputer i człowiek uzyskają różne wyniki, to lepiej postawić wszystkie pieniądze na to, że to komputer ma rację. Nie jest to jednak przesądzone. Komputer wiernie realizujący sprawdzony program również może się pomylić – na przykład jakiś promień kosmiczny może uderzyć w komórkę jego pamięci i zmienić 0 na 1. Można się przed tym zabezpieczyć, powtarzając obliczenia dwukrotnie, ale jeszcze większe ryzyko wystąpienia błędów wiąże się z tym, że również projektanci układów scalonych mogą się mylić. Procesor Intel Pentium P5 miał błąd w procedurze obliczeń zmiennoprzecinkowych – jeśli kazało mu się podzielić 4195835 przez 3145727, procesor dawał wynik 1,33373, podczas gdy poprawna

wartość to 1,33382. Jak się okazało, źródłem błędu był brak kilku pozycji w tablicy używanej przy dzieleniu[35]. Również inne elementy mogą doprowadzić do błędnych wyników, na przykład system operacyjny komputera lub błędy w samym programie.

Wylano już całe morze atramentu na snucie filozoficznych rozważań na temat tego, jak wspomagany komputerowo dowód Appela i Hakena zmienił samą naturę „dowodzenia". Rozumiem, do czego zmierzają filozofowie, ale pojęcie dowodu, jakim matematycy posługują się w praktyce, nie jest tym samym, którego naucza się na zajęciach z logiki. A nawet jeśli zgodzimy się, że obowiązuje takie bardziej formalne ujęcie, to nigdzie nie zapisano, że każdy krok wywodu logicznego musi być sprawdzony przez człowieka. Matematycy już od stuleci wykorzystują różne urządzenia do przeprowadzania rutynowych obliczeń. Ba, nawet jeśli ktoś sprawdzi dany dowód linijka po linijce i nie znajdzie w nim błędów, to jaką możemy mieć pewność, że czegoś nie przeoczył? Doskonała, niepodważalna logika jest ideałem, do którego dążymy. Jako istoty niedoskonałe robimy wszystko, co w naszej mocy, żeby się do niego zbliżyć, ale nigdy nie uda nam się pozbyć wszystkich wątpliwości.

W książce *Four Colours Suffice* (*Wystarczą cztery kolory*) Robin Wilson trafnie wskazał główny socjologiczny aspekt takiej reakcji społeczności uczonych:

> Naukowcy podzielili się na dwie grupy: uczonych po czterdziestce nie sposób było przekonać, że dowód komputerowy może być poprawny, natomiast ci, którzy nie ukończyli jeszcze czterdziestu lat, nie przyjmowali do wiadomości, że poprawny może być dowód obejmujący 700 stron ręcznych obliczeń.

Jeśli skonstruowane przez nas urządzenia są od nas lepsze w jakichś zadaniach, to warto z nich korzystać. *Techniki* dowodzenia mogą ulegać zmianie – tak się zresztą cały czas dzieje, bo na tym polega natura badań naukowych. Natomiast samo pojęcie dowodu nie zmienia się radykalnie, jeśli niektóre kroki wykona komputer.

Dowód jest opowieścią. Dowód wspomagany komputerowo jest opowieścią zbyt długą, by ją można było prześledzić w całości, musimy więc zadowolić się zwięzłym streszczeniem i opasłym dodatkiem zawierającym szczegóły obliczeń komputerowych.

Od czasu pionierskich prac Appela i Hakena matematycy przyzwyczaili się już do korzystania ze wspomagania komputerowego. Wciąż *wyżej cenią* dowody przeprowadzone w całości przez człowieka, ale nie traktują tego już jako wymagania. W latach dziewięćdziesiątych dowód Appela i Hakena wciąż jednak otaczała uzasadniona aura niepewności. Zamiast więc jedynie sprawdzić ich pracę, niektórzy matematycy postanowili powtórzyć cały dowód, wykorzystując do tego nowe rozwiązania techniczne i dużo lepsze komputery. W 1994 roku Neil Robertson, Daniel Sanders, Paul Seymour i Robin Thomas pominęli wszystko, co znalazło się w artykule Appela i Hakena, z wyjątkiem podstawowej strategii. W ciągu roku odkryli nieunikniony zbiór 633 konfiguracji, których redukowalność można było udowodnić, używając jedynie 32 reguł rozładowywania. Takie rozwiązanie jest dużo prostsze od wyniku Appela i Hakena, który obejmował 1482 konfiguracje i 487 reguł rozładowywania. Współczesne komputery są tak szybkie, że cały dowód można sprawdzić na domowym laptopie w zaledwie kilka godzin.

No świetnie, ale komputer wciąż jest górą. Czy możemy się go pozbyć? Mamy coraz silniejsze poczucie, że w przypadku tej konkretnej opowieści nie jest wcale wykluczone, iż ludziom uda się w końcu objąć umysłem cały dowód. Być może jakieś nowe idee związane z problemem czterech barw doprowadzą do uzyskania prostszego dowodu, wymagającego niewielkiego tylko wsparcia komputerowego (a może obywającego się w ogóle bez takiej pomocy), a wtedy matematycy będą mogli go po prostu przeczytać, pomyśleć chwilę i stwierdzić: „Tak, zgadza się!”. Nie mamy jeszcze takiego dowodu i niewykluczone, że on w ogóle nie istnieje, ale coś wisi w powietrzu…

Matematycy dowiadują się ciągle wielu nowych rzeczy na temat sieci. Specjaliści od topologii i geometrii odkrywają głębokie związki między sieciami i zupełnie innymi działami matematyki – niektóre z tych odkryć mają nawet zastosowanie w fizyce matematycznej. Jednym z pojęć, które się od czasu do czasu przewijają w tych badaniach, jest pojęcie krzywizny. Jego nazwa mówi praktycznie wszystko: krzywizna przestrzeni jest miarą tego, jak bardzo się ona wykrzywia. Gdy przestrzeń jest płaska, tak jak na przykład płaszczyzna, jej krzywizna jest zerowa. Jeśli zakrzywia się tak samo we wszystkich kierunkach, niczym szczyt górski, który wszędzie jest wygięty w dół, to ma krzywiznę dodatnią. Jeśli natomiast przypomina górską przełęcz i zakrzywia się w górę w niektórych kierunkach, a w dół w innych, to ma krzywiznę ujemną. Istnieją twierdzenia geometryczne wywodzące się ze wzoru Eulera, które pozwalają powiązać sieci narysowane w danej przestrzeni z jej krzywizną. O takiej możliwości świadczył już wzór Heawooda dla torusa z określoną liczbą otworów. Sfera ma dodatnią krzywiznę, torus można przedstawić jako kwadrat, w którym utożsamiono ze sobą przeciwległe boki (zob. ryc. 12, po prawej), a zatem ma krzywiznę zerową, natomiast torus z dwoma otworami lub większą ich liczbą ma krzywiznę ujemną. Istnieje więc jakiś związek między krzywizną a kolorowaniem mapy.

Podstawą tego związku jest pewna użyteczna cecha krzywizny – to mianowicie, że trudno się jej pozbyć. Jest niczym kot pod dywanem. Jeśli dywan leży płasko, to nie ma pod nim kota, ale jeśli jest gdzieś wybrzuszony, to możemy być pewni, że znajdziemy tam kota. Możemy go przeganiać w różne strony, ale w ten sposób zdołamy jedynie przesunąć wybrzuszenie gdzieś indziej. Podobnie krzywiznę można przesunąć, ale nie – usunąć. No chyba że kot dotrze do brzegu dywanu i uda mu się uciec, zabierając ze sobą swoją krzywiznę. Opracowane przez Heescha reguły rozładowywania przypominają trochę zakamuflowaną krzywiznę. Pozwalają na przesuwanie ładunku elektrycznego w różne miejsca, ale go nie niszczą. Czy może istnieć jakieś pojęcie krzywizny sieci i jakaś sprytna reguła rozładowywania, która w istocie przesuwa krzywiznę?

Jeśli tak, to może udałoby się nakłonić sieć do tego, żeby sama się automatycznie pokolorowała. Wystarczyłoby w tym celu przypisać jej wierzchołkom (a może i krawędziom) krzywiznę i pozwolić, żeby sieć sama rozłożyła ją bardziej równomiernie. Może określenie „równomiernie" implikuje w tym wypadku, że wystarczą tylko cztery kolory, jeśli wszystko się odpowiednio ustawi? To tylko jeden z pomysłów, zresztą nie mój, i nie wyjaśniliśmy go tu na tyle szczegółowo, żeby mógł nabrać większego sensu. Pozwala nam jednak uzmysłowić sobie, jak działa intuicja matematyczna, i daje nadzieję, że w przyszłości uda się być może znaleźć jakiś bardziej konceptualny dowód twierdzenia o czterech barwach – taki, który będzie przypominał wątek prowadzący do celu, a nie streszczenie z książką telefoniczną w roli dodatku. Z podobną ideą spotkamy się w dużo bardziej złożonym kontekście w rozdziale 10 i przekonamy się, że takie podejście pozwoliło rozwiązać jeszcze większy problem z dziedziny topologii.

28 Louis H. Kauffman, *Map coloring and the vector cross product* (*Kolorowanie map a iloczyn wektorowy*), „Journal of Combinatorial Theory B" 1990, tom 48, s. 145–154. Louis H. Kauffman, *Reformulating the map color theorem* (*Nowe sformułowanie twierdzenia o kolorowaniu map*), „Discrete Mathematics" 2005, tom 302, s. 145–172.

29 Jeśli dopuścimy, żeby granice były naprawdę skomplikowane, dużo bardziej powyginane od tych, jakie spotyka się na mapach, to wówczas dowolna liczba krajów może mieć wspólną „granicę". Konstrukcja pod nazwą „jeziora Wady" dowodzi, że takie sprzeczne z intuicją twierdzenie jest prawdziwe (zob. http://en.wikipedia.org/wiki/Lakes_of_Wada).

30 Ściśle rzecz biorąc, powinniśmy mówić o „grafie dualnym", ponieważ tradycyjnie matematycy posługiwali się „grafami", a nie „sieciami". Obecnie jednak określenie „sieć" zyskuje coraz bardziej na popularności, budzi szersze skojarzenia i pozwala uniknąć pomyłek związanych z innymi znaczeniami słowa „graf".

31 Przez długi czas sądzono, że artykuł z „Nature" był ostatnim nawiązaniem do tego problemu w czasopismach naukowych przed niemal stuletnią przerwą, ale historyk matematyki Robin Wilson odnalazł niedawno ten późniejszy artykuł Cayleya.

32 Przyjmijmy, że *S* jest liczbą ścian sieci dualnej (z uwzględnieniem jednej dużej ściany otaczającej całą sieć), *K* liczbą krawędzi, *W* zaś liczbą węzłów. Możemy założyć, że każda ściana sieci dualnej ma przynajmniej trzy krawędzie – jeśli występuje w niej ściana z dwiema krawędziami, to odpowiada ona „nadmiarowemu" wierzchołkowi w sieci wyjściowej, połączonemu tylko dwiema krawędziami. Taki wierzchołek można usunąć, a obie krawędzie – połączyć. Każda krawędź oddziela od siebie dwie ściany, a każda ściana ma przynajmniej trzy krawędzie, zatem $K \geq 3S/2$ lub równoważnie: $2K/3 \geq S$. Z twierdzenia Eulera wiemy, że $S + W - K = 2$, a zatem $2K/3 + W - K \geq 2$, z czego wynika, że

$$12 + 2K \leq 6W$$

Przyjmijmy, że W_m jest liczbą wierzchołków mających *m* sąsiadów. Nas interesuje przypadek, gdy W_2, W_3, W_4 oraz W_5 są równe zero, a wtedy:

$$W = W_6 + W_7 + W_8 + \cdots$$

Ponieważ każda krawędź łączy dwa wierzchołki, mamy

$$2K = 6W_6 + 7W_7 + 8W_8 + \cdots$$

Wstawiwszy to do nierówności, otrzymujemy:

$$12 + 6W_6 + 7W_7 + 8W_8 + \cdots \leq 6W_6 + 6W_7 + 6W_8 + \cdots$$

wtedy

$$12 + W_7 + 2W_8 + \cdots \leq 0$$

a to jest niemożliwe

33 Określenie „łańcuch" jest w tym wypadku nieco mylące, ponieważ sugeruje, że mamy do czynienia z liniowym ciągiem elementów. Łańcuch Kempe'a może zawierać pętle, a nawet się rozgałęziać.

34 Pełny dowód można znaleźć w książce Gerharda Ringela *Map Color Theorem* (*Twierdzenie o kolorowaniu map*). Rozpatruje się w nim 12 przypadków, zależnych od tego, jaką postać ma genus: $12k$, $12k + 1$, …, $12k + 11$. Oznaczmy te przypadki liczbami od 0 do 11. Poza skończoną liczbą wyjątków przypadki te rozwiązali następujący uczeni:

- przypadek 5: Ringel w 1954 roku,
- przypadki 3, 7 oraz 10: Ringel w 1961 roku,
- przypadki 0 oraz 4: Charles M. Terry, Lloyd Welch i John W.T. Youngs w 1963 roku,
- przypadek 1: William Gustin i John W.T. Youngs w 1964 roku,
- przypadek 9: Gustin w 1965 roku,
- przypadek 6: Youngs w 1966 roku,
- przypadki 2, 8 oraz 11: Ringel i Youngs w 1967 roku.

Wspomniane wyjątki to przypadki o genusie 18, 20, 23 (rozwiązał je Yves Mayer w 1967 roku) oraz 30, 35, 47 i 659 (rozwiązane przez Ringela i Youngsa w 1968 roku). Uczeni uporali się również z analogicznym problemem dla powierzchni jednostronnych (takich jak wstęga Möbiusa, ale bez brzegu), którymi Percy Heawood również się zajmował.

35 Niezwykłą historię odkrycia tego błędu i dalszych związanych z tym wydarzeń można przeczytać (w języku angielskim) na stronie internetowej: http://en.wikipedia.org/wiki/Pentium_FDIV_bug.

5. Symetria doskonała

Hipoteza Keplera

Wszystko zaczęło się od płatka śniegu.

Śnieg ma nieodparty urok. Spada z nieba w postaci puszystych białych płatków, podmuchy wiatru usypują z niego miękkie pagórki i wzgórza, które pokrywają ziemię aż po horyzont i bez niczyjej pomocy zmieniają się w dziwaczne kształty. Jest zimny. Można po nim jeździć na nartach i sankach, lepić z niego śnieżki i bałwany… ale potrafi też być niebezpieczny: czasami wystarczy odrobina nieuwagi, żeby zasypały nas tysiące ton śniegu. Gdy w końcu znika, nie unosi się z powrotem w kierunku nieba – a przynajmniej nie pod postacią białych płatków. Przekształca się w zwyczajną wodę, która może oczywiście wyparować i unieść się w powietrze, ale znaczna jej część spływa do rzek, a nimi dalej, do mórz i oceanów. Śnieg jest jedną z postaci lodu, a lód to zamarznięta woda.

Żadna to nowina. Wszystko to musiało być oczywiste już dla neandertalczyków.

Płatki śniegu nie są bezkształtnymi grudkami. Świeże płatki, zanim jeszcze zaczną się topić, wyglądają często jak maleńkie, filigranowe gwiazdki – są płaskie, sześcioramienne i symetryczne. Inne przypominają zwyczajne sześciokąty. Niektóre są mniej symetryczne lub rozrastają się w trzech wymiarach, ale sześcioramienne płatki śniegu są najbardziej charakterystyczne i najczęściej spotykane. Płatki śniegu są kryształkami lodu. To również nic nowego: wystarczy spojrzeć na płatek, żeby od razu zauważyć, że jest kryształem. Nie są to jednak zwyczajne kryształy o płaskich ścianach w kształcie wielokątów. Ich najbardziej zagadkowa cecha wynika ze szczypty

chaosu – mimo że wszystkie mają taką samą symetrię, różnią się między sobą pod względem szczegółów budowy. Powiada się, że nie ma dwóch takich samych płatków śniegu. Zawsze ciekawiło mnie, skąd to wiadomo, ale liczby potwierdzają to stwierdzenie, pod warunkiem że przyjmiemy wystarczająco drobiazgową definicję określenia „takie same".

Dlaczego płatki śniegu są sześcioboczne? Czterysta lat temu jeden z największych matematyków i astronomów XVII stulecia postawił sobie właśnie takie pytanie i zaczął na ten temat rozmyślać. Doszedł do zdumiewająco trafnej odpowiedzi, która zadziwia tym bardziej, że nie przeprowadził żadnego doświadczenia. Po prostu zestawił ze sobą kilka prostych idei znanych każdemu człowiekowi. Poskładał je w całość niczym nasiona granatu upakowane we wnętrzu owocu.

Nazywał się Johannes Kepler i zaczął rozmyślać o płatkach śniegu z bardzo ważnego powodu. Pod względem materialnym był całkowicie uzależniony od swojego mecenasa Johannesa Wackhera von Wackhenfelds. W tamtym okresie Kepler był nadwornym matematykiem cesarza rzymskiego Rudolfa II, Wackher zaś dyplomatą i doradcą samego cesarza. Kepler chciał sprawić swojemu mecenasowi jakiś prezent z okazji Nowego Roku. Najlepiej, gdyby było to coś niedrogiego, niezwykłego i intrygującego. Powinno być to coś, dzięki czemu Wackher zrozumie, jak wspaniałych odkryć udaje się dokonać dzięki jego wsparciu finansowemu. Kepler zebrał więc swoje przemyślenia na temat płatków śniegu w niewielkiej książeczce i podarował ją Wackherowi jako prezent noworoczny. Jej tytuł brzmi *Noworoczny podarek albo O sześciokątnych płatkach śniegu*[36]. Był rok 1611. Gdzieś w środku książki Kepler zapisał w postaci krótkiej uwagi jeden z ważniejszych kroków w swoim wywodzie – uwaga ta okazała się zagadką matematyczną, której nie udawało się rozwiązać przez 387 lat.

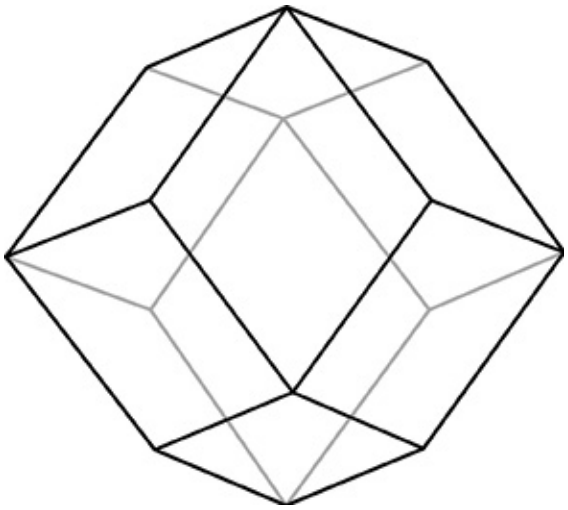
Kepler był niestrudzonym poszukiwaczem wzorów. Jego najważniejszym osiągnięciem naukowym było odkrycie trzech podstawowych praw ruchu planet, z których pierwsze i najbardziej znane głosi, że orbity mają kształt elips. Był również mistykiem

całkowicie zanurzonym w pitagorejskim obrazie świata, zgodnie z którym Wszechświat opiera się na liczbach, wzorach i figurach matematycznych. Oprócz astronomii zajmował się również astrologią: w tamtych czasach matematycy często dorabiali jako astrologowie, ponieważ potrafili przeprowadzić obliczenia pozwalające ustalić, czy Strzelec jest w ascendencie. Zamożni mecenasi, nawet członkowie rodzin królewskich, chętnie płacili za postawienie horoskopu.

W swojej książce Kepler zwraca uwagę na fakt, że śnieg jest na początku parą wodną, która jest przecież bezkształtna, a mimo to w jakiś sposób zmienia się ona w sześcioboczne sztywne płatki. Tę zmianę musi wywoływać jakiś czynnik, twierdził Kepler:

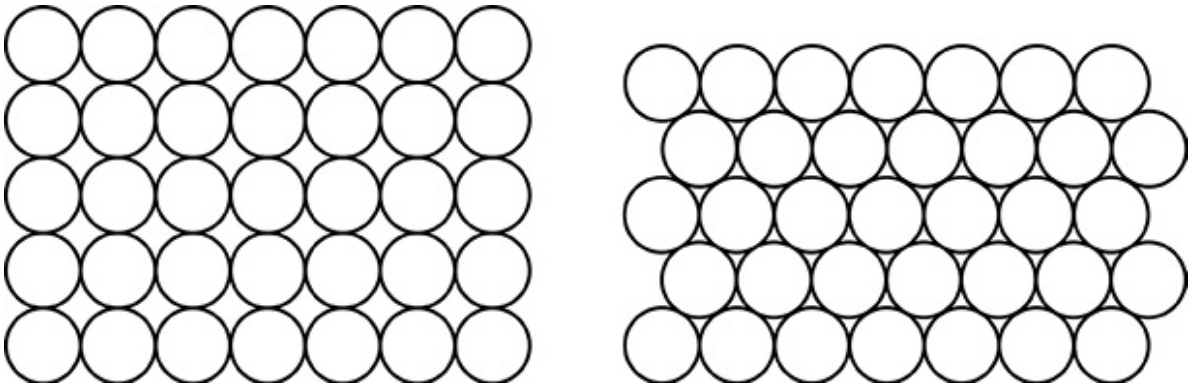
> Czy z konieczności materii [czynnik ów] tworzy kształt sześciokątny, czy ze swej natury, dla której wrodzone są pierwotny wzór piękna zawarty w sześciokącie albo znajomość celu, do którego ten kształt prowadzi?[37]

Poszukując odpowiedzi, rozważał również inne przykłady występowania sześciokątów w naturze. Jeśli się nad tym zastanowimy, od razu przyjdą nam na myśl plastry miodu tworzone przez pszczoły. Są one zbudowane z dwóch warstw sześciokątnych komórek, umieszczonych jedna za drugą, tak że ściany, którymi się stykają, mają kształt trzech rombów – równoległoboków o równych bokach. Kształt ten przypomniał Keplerowi bryłę o nazwie dwunastościan rombowy (zob. ryc. 15). Nie jest to jeden z pięciu wielościanów foremnych opisanych przez Euklidesa i znanych już pitagorejczykom, ale mimo to bryła ta ma pewną charakterystyczną własność: jednakowymi dwunastościanami rombowymi można wypełnić całą przestrzeń, tak że nie zostaną żadne luki. Ten sam kształt występuje również w owocach granatowca, w których niewielkie okrągłe nasionka rozrastają się w dużym ścisku, co sprawia, że muszą się upakować w jak najefektywniejszy sposób.



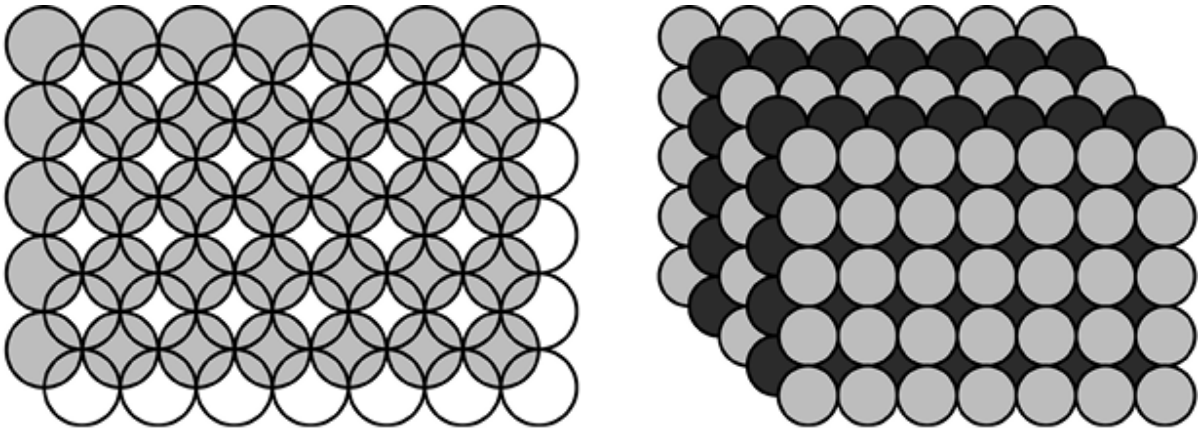
Ryc. 15. Dwunastościan rombowy – bryła, której ściany tworzy 12 rombów

Jak każdy szanujący się matematyk, Kepler wyszedł od najprostszego przypadku kul ułożonych w jedną, płaską warstwę. Jest to równoważne upakowaniu jednakowych kół na płaszczyźnie. W takiej sytuacji istnieją tylko dwa naturalne układy. Pierwszy polega na ułożeniu kul w rogach kwadratów (zob. ryc. 16, po lewej), drugi zaś – na umieszczeniu ich w wierzchołkach trójkątów równobocznych (zob. ryc. 16, po prawej). Układy takie, powtórzone na całej nieskończonej płaszczyźnie, tworzą, odpowiednio, sieć kwadratową i sieć trójkątną. Słowo „sieć" odnosi się do ich przestrzennego rozmieszczenia, które ma charakter okresowy, czyli powtarzającego się w obydwu niezależnych kierunkach. Na rysunkach można oczywiście pokazać tylko ograniczony wycinek całego wzoru, nie należy więc zwracać uwagi na brzegi tych obszarów. To samo zastrzeżenie dotyczy rycin od 17 do 20. Po obu stronach na rycinie 16 pokazano pięć rzędów kul ułożonych w taki sposób, że w każdym rzędzie kule stykają się ze wszystkimi sąsiadkami. Zauważmy jednak, że sieć trójkątna jest nieco bardziej ściśnięta – jej rzędy znajdują się bliżej siebie. Wynika z tego, że w sieci trójkątnej kule są ciaśniej upakowane niż w sieci kwadratowej.



Ryc. 16. Po lewej: Sieć kwadratowa. Po prawej: Sieć trójkątna (nazywana również sześciokątną).

Następnie Kepler zastanowił się nad tym, jak mogą być umieszczone względem siebie kolejne warstwy, i wyróżnił cztery przypadki. W pierwszych dwóch wszystkie warstwy są utworzone z sieci kwadratowych. Jeden z możliwych sposobów polega na umieszczeniu kolejnych warstw w taki sposób, by kule w sąsiednich warstwach znajdowały się bezpośrednio jedna nad drugą. W takiej sytuacji każda kula ma sześć bezpośrednich sąsiadek: cztery we własnej warstwie, jedną u góry i jedną na dole. Takie upakowanie przypomina trójwymiarową szachownicę skonstruowaną z sześcianów – właśnie taką szachownicę uzyskalibyśmy, gdybyśmy nadmuchali wszystkie kule do tego stopnia, że nie miałyby się już gdzie rozszerzać. To jednak – stwierdza Kepler – „nie jest to najgęstsze ułożenie”[38]. Kule upakujemy jeszcze ściślej, jeśli przesuniemy drugą warstwę w bok, tak by jej kule wpasowały się w puste miejsca między kulami w dolnej warstwie (zob. ryc. 17, po lewej). Następnie należy postąpić tak samo z kolejnymi warstwami (zob. ryc. 17, po prawej). Teraz każda kula ma dwanaście sąsiadek: cztery we własnej warstwie, cztery u góry i cztery na dole. Gdybyśmy je nadmuchali, to wypełnilibyśmy przestrzeń dwunastościanami rombowymi.



Ryc. 17. Po lewej: Dodanie drugiej warstwy kul (puste kółka) położonej na pierwszej (kółka w kolorze szarym). Po prawej: Ten sam proces powtórzony wielokrotnie.

W pozostałych dwóch przypadkach warstwy są sieciami trójkątnymi. Jeśli umieścimy je jedna na drugiej w taki sposób, że sąsiednie kule znajdą się dokładnie jedna nad drugą, to wtedy każda kula będzie miała osiem sąsiadek: sześć we własnej warstwie, jedną u góry i jedną na dole. Kolejne warstwy można też umieścić tak, by kule następnej warstwy wypełniły pustą przestrzeń między kulami w warstwie dolnej. Teraz każda kula ma dwanaście sąsiadek: sześć we własnej warstwie, trzy na górze i trzy na dole. Jest to taka sama liczba sąsiadek jak w drugim układzie warstw kwadratowych i Kepler przedstawia szczegółową analizę geometrii tych układów, z której wynika, że ten czwarty układ jest w istocie taki sam jak drugi. Jedyna różnica polega na tym, że warstwy kwadratowe nie są ułożone poziomo, ale nachylone pod pewnym kątem. „Tak więc – pisze – w najgęstszym ułożeniu w przestrzeni nie może istnieć wzór trójkątny bez czworokątnego i odwrotnie”[39]. Jeszcze powrócimy do tego stwierdzenia – jest ono bardzo ważne.

Uporawszy się z podstawową geometrią upakowania kul, Kepler powraca do płatka śniegu i jego sześciokrotnej symetrii. Widzi w nim podobieństwo do upakowania kul na płaszczyźnie w trójkątną sieć w taki sposób, że każda kula jest otoczona sześcioma innymi,

ułożonymi w kształt sześcianu foremnego. Właśnie to musi tłumaczyć, dlaczego płatki śniegu są sześciokątne – konkluduje.

Ten rozdział nie jest poświęcony płatkom śniegu, ale wyjaśnienie ich symetrii zaproponowane przez Keplera jest bardzo podobne do tego, jakie przedstawilibyśmy dzisiaj, nie zaszkodzi więc zatrzymać się na chwilę przy tej kwestii. Jak to możliwe, że są one – *mogą* być – tak różnorodne, a przy tym symetryczne? Gdy woda ulega krystalizacji i powstaje lód, atomy wodoru i tlenu tworzące cząsteczki wody zbijają się w symetryczną strukturę – w sieć krystaliczną. Sieć ta jest bardziej złożona niż układy kul, które rozważał Kepler, ale jej dominującą cechą jest sześciokrotna symetria. Płatek śniegu rozrasta się z maleńkiego „zarodka", który tworzy kilka zaledwie atomów ułożonych w niewielki fragment sieci. Zarodek ten ma taką samą sześciokrotną symetrię i właśnie to decyduje o dalszym wzroście kryształka lodu, który rozrasta się zgodnie z kierunkiem podmuchów wiatrów wiejących we wszystkie strony we wnętrzu chmury burzowej.

Wielka różnorodność wzorów płatków śniegu wynika ze zmiennych warunków panujących w chmurze. W zależności od temperatury i wilgotności kryształy mogą rozrastać się albo jednorodnie, a wtedy na całej powierzchni kryształu nowe atomy dodawane są w takim samym tempie i powstają równe sześciany, albo dendrytycznie, a wtedy tempo wzrostu w każdym miejscu jest inne i powstaje struktura drzewiasta. Rozrastający się płatek śniegu przemieszcza się w górę i w dół chmury, co powoduje, że warunki wokół niego zmieniają się w sposób przypadkowy. Płatek jest jednak tak mały, że w każdej chwili warunki te są w zasadzie takie same we wszystkich jego sześciu rogach. Zatem we wszystkich rogach proces wzrostu przebiega tak samo. Każdy płatek śniegu jest zapisem jego historii. W rzeczywistości jego sześciokrotna symetria nigdy nie jest dokładna, ale bardzo często odchylenia są znikome. Lód jest dziwną substancją, która może tworzyć również inne kształty – spiczaste sople, płaskie tafle, sześciokątne graniastosłupy czy graniastosłupy z taflami na górze. Pełne wyjaśnienie tych procesów jest bardzo złożone, ale wszystko wynika z tego, jak ułożone są atomy

w krysztale lodu[40]. W czasach Keplera teoria atomowa była w najlepszym razie mglistą sugestią wysuniętą przez starożytnych Greków, tym bardziej więc zdumiewa, jak wiele udało się uczonemu wyjaśnić, skoro miał do dyspozycji tylko proste obserwacje, eksperymenty myślowe i umiejętność dostrzegania wzorców.

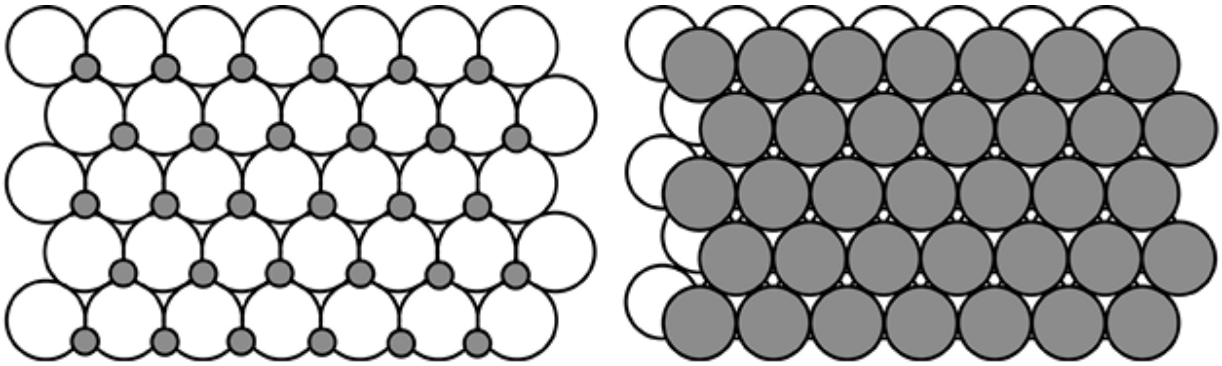
Hipoteza Keplera nie dotyczy samych płatków śniegu. Jest nią owo zamieszczone w ramach dygresji spostrzeżenie, że ułożenie wielu warstw ściśle upakowanych kul w taki sposób, że warstwa wyższa wypełnia luki w niższej, prowadzi do „najgęstszego ułożenia w trzech wymiarach". Można to wyrazić w sposób mniej ścisły: jeśli chcemy zapakować dużą liczbę pomarańczy do dużego pudła, maksymalnie je zapełniając, to powinniśmy ułożyć owoce tak, jak zrobiłby to każdy doświadczony sprzedawca z warzywniaka.

Trudność nie polega w tym wypadku na znalezieniu *odpowiedzi*. Kepler powiedział nam, jak ona brzmi. Trudność polega na udowodnieniu, że miał rację. W ciągu następnych stuleci uczeni zgromadzili wiele dowodów pośrednich. Nikomu nie udało się uzyskać gęstszego upakowania. Taki sam układ atomów występuje w wielu kryształach, w których dobre upakowanie wynika zapewne z dążenia do minimalizacji energii – jest to jedna z podstawowych zasad decydujących o powstaniu wielu kształtów występujących w przyrodzie. Tego rodzaju dowody były zadowalające dla większości fizyków. Równocześnie nikomu nie udało się przedstawić dowodu, że lepszy sposób upakowania *nie istnieje*. Okazało się, że prostsze pytania tego samego rodzaju, na przykład kwestia upakowania kół na płaszczyźnie, mają ukrytą głębię. Cały ten obszar badań jest trudny i pełen niespodzianek. Wszystko to martwiło matematyków, mimo że większość z nich była przekonana, iż Kepler podał poprawną odpowiedź. W 1958 roku Claude Ambrose Rogers nazwał hipotezę Keplera czymś, w co „wielu matematyków wierzy, a wszyscy fizycy to wiedzą"[41]. W tym rozdziale opowiemy, jak matematycy zmienili tę wiarę w pewność.

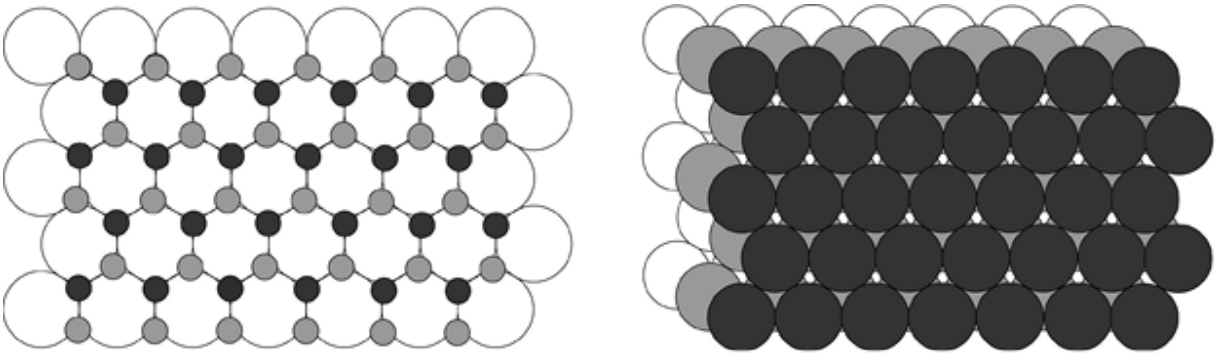
Aby zrozumieć, czego dokonali, musimy najpierw przyjrzeć się uważnie Keplerowskiemu układowi kul, który fachowo nazywa się

układem regularnym (lub sześciennym) centrowanym ściennie. Gdy to zrobimy, zaczniemy dostrzegać subtelności związane z tym problemem. Pierwszym pytaniem, jakie ciśnie się na usta, jest: dlaczego używamy warstw kwadratowych. W końcu najciaśniejsze upakowanie pojedynczej warstwy występuje w przypadku sieci *trójkątnej*. Odpowiedzią jest stwierdzenie, że układ regularny centrowany ściennie można uzyskać również z warstw trójkątnych. Właśnie to miał na myśli Kepler, gdy mówił, że „wzór trójkątny nie może istnieć bez czworokątnego". Łatwiej jednak można opisać taką sieć przy użyciu warstw kwadratowych, a dodatkowo przy okazji zrozumiemy, że hipoteza Keplera nie jest tak oczywista jak pakowanie pomarańczy w sklepie warzywnym.

Wyobraźmy sobie najpierw płaską warstwę kul ułożonych w trójkąty (zob. ryc. 16, po prawej). Między kulami znajdują się luki w kształcie zakrzywionych trójkątów, w których możemy zmieścić następną warstwę kul. Gdy wychodziliśmy od warstwy kwadratowej, mogliśmy wykorzystać wszystkie luki, zatem umiejscowienie drugiej warstwy – i następnych – było jednoznacznie określone. Sytuacja jest inna, gdy zaczynamy od układu trójkątnego. Nie możemy wykorzystać wszystkich luk, ponieważ znajdują się one zbyt blisko siebie. Zdołamy zapełnić tylko połowę z nich. Jeden z możliwych układów zamieszczono na rycinie 18 (po lewej), używając dla przejrzystości małych szarych kropek, a po prawej stronie tej ryciny pokazano, jak powinna być wówczas umieszczona następna warstwa kul. Drugi sposób umieszczenia nowej warstwy w lukach warstwy pierwszej pokazano na rycinie 19 (po lewej), używając ciemniejszych kropek. Kropki te zgadzają się z lukami w warstwie drugiej, dodajemy więc w tych miejscach warstwę trzecią – wynik tej operacji zaprezentowano na rycinie 19 (po prawej).



Ryc. 18. Jeden ze sposobów wpasowania sieci trójkątnej w luki w poprzedniej warstwie

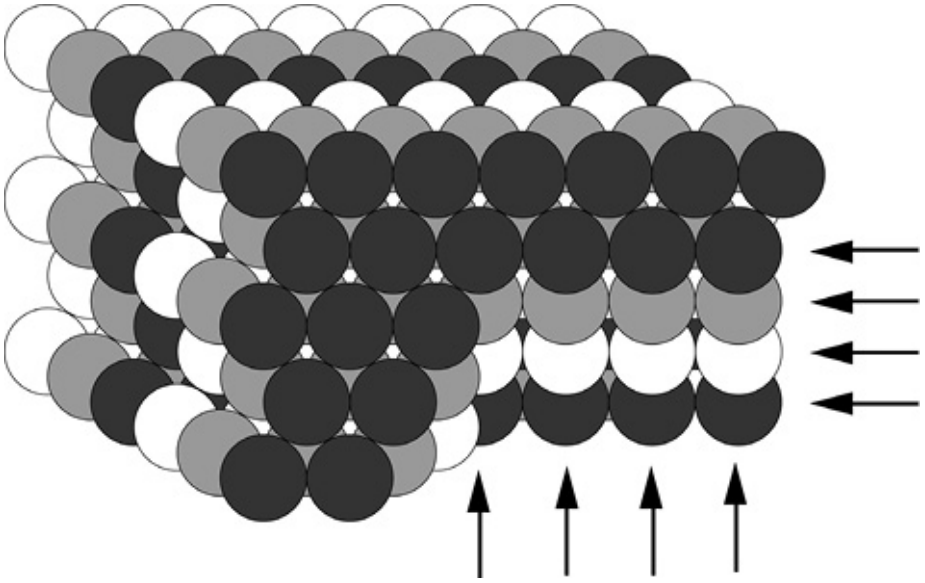


Ryc. 19. Sposoby umieszczenia kolejnych warstw sieci trójkątnej

Różnica między tymi układami nie ma w istocie znaczenia, gdy rozważamy tylko dwie warstwy. Jeśli obrócimy drugi układ o 60 stopni, dostaniemy pierwszy. Układy te są takie same „z dokładnością do symetrii". Jednak po ułożeniu pierwszych dwóch warstw mamy dwie, naprawdę różne możliwości umieszczenia warstwy trzeciej. Każda nowa warstwa ma dwa układy luk, pokazane na rycinie 19 (po lewej) za pomocą jasnoszarych i ciemnoszarych kropek. Jeden z tych układów zgadza się ze środkami kul w warstwie położonej bezpośrednio pod ostatnią warstwą – na rycinie 19 (po prawej) miejsca te widać jako jasnoszare trójkąty widoczne między ciemnoszarymi kółkami. Drugi układ natomiast zgadza się z lukami w warstwie położonej jeszcze jeden poziom niżej – na rycinie 19 (po prawej) miejsca te są widoczne jako trójkąty z małymi białymi

sześciokątami w środku. Aby uzyskać układ regularny centrowany ściennie, musimy umieścić trzecią warstwę w punktach oznaczonych ciemnymi kropkami, a potem dalej powtarzać ten układ warstw w nieskończoność.

Nie jest wcale oczywiste, że otrzymamy w ten sposób układ regularny centrowany ściennie. Czy gdzieś tu są kwadraty? Ależ tak, są, ale należy ich szukać w płaszczyznach nachylonych pod pewnym kątem. Na rycinie 20 przedstawiono sześć kolejnych warstw trójkątnych, z których usunięto kilka kul. Strzałki pokazują rzędy i kolumny sieci kwadratowej, która ukrywa się w środku. Warstwy równoległe do tej pokazanej również są sieciami kwadratowymi umieszczonymi jedna na drugiej w taki sam sposób, w jaki skonstruowaliśmy wcześniej układ regularny centrowany ściennie.



Ryc. 20. Wewnątrz warstw trójkątnych ukrywają się warstwy kwadratowe nachylone pod pewnym kątem.

Jak „ścisłe” jest to upakowanie? Ścisłość (skuteczność, ciasność) upakowania mierzymy, podając jego gęstość, czyli wartość mówiącą o tym, jaki fragment całej przestrzeni zajmują kule[42]. Im większa gęstość, tym gęstsze upakowanie. Sześciany można upakować z gęstością wynoszącą 1, wypełniają więc całą przestrzeń. Kule

zostawiają oczywiście luki, zatem gęstość będzie mniejsza od 1. Dla układu regularnego centrowanego ściennie gęstość wynosi dokładnie $\pi/\sqrt{18}$, czyli w przybliżeniu 0,7405. Zatem przy takim upakowaniu kule wypełniają niecałe trzy czwarte przestrzeni. Hipoteza Keplera głosi, że żadne inne upakowanie kul nie będzie miało większej gęstości.

Sformułowaliśmy tę hipotezę dość ostrożnie. Nie powiedzieliśmy: „upakowanie regularne centrowane ściennie ma największą możliwą gęstość". To nieprawda, o czym można się łatwo przekonać. Wystarczy wrócić do konstrukcji układu regularnego centrowanego ściennie z wykorzystaniem warstw trójkątnych. Powiedzieliśmy wtedy, że po położeniu dwóch pierwszych warstw trzecią można ułożyć na dwa różne sposoby. Układ regularny centrowany ściennie powstanie wtedy, gdy wybierzemy drugie ułożenie, zaznaczone ciemnoszarymi kropkami. A co się stanie, jeśli wybierzemy ustawienie zaznaczone na rysunku jasnoszarymi kropkami? Wtedy warstwa 3 znajdzie się dokładnie nad warstwą 1. Jeśli będziemy dalej postępowali tak samo, układając każdą nową warstwę dokładnie nad położoną dwa poziomy niżej, to otrzymany inny rodzaj sieci – sieć o układzie heksagonalnym. Jest to układ zupełnie inny od upakowania regularnego centrowanego ściennie, ma jednak taką samą gęstość. Zgadza się to z naszą intuicją, ponieważ oba różne sposoby umieszczenia trzeciej warstwy są związane ze sobą symetrią obrotową, muszą więc równie dobrze wpasowywać się w warstwę poprzednią.

To jedyne upakowania *o charakterze sieciowym*, które można otrzymać przez układanie kolejnych warstw trójkątnych, ale w 1883 roku geolog i krystalograf William Barlow zwrócił uwagę na fakt, że możemy również określać położenie kolejnych warstw w sposób losowy, wybierając za każdym razem jedną z tych dwóch możliwości. Ponieważ oba ułożenia wnoszą taki sam wkład do gęstości, to wszystkie upakowania będą miały gęstość $\pi/\sqrt{18}$. Istnieje nieskończenie wiele takich losowych ciągów, mamy więc nieskończenie wiele różnych upakowań o takiej samej gęstości.

Jednym słowem, nie istnieje jedno najgęstsze upakowanie kul. Mamy wiele takich upakowań i wszystkie cechuje taka sama gęstość. Ten brak jednoznaczności stanowi ostrzeżenie: problem ten nie jest tak prosty, jak się wydaje. Największa *gęstość* jest tylko jedna, jeśli Kepler miał rację, ale istnieje nieskończenie wiele układów charakteryzujących się taką właśnie gęstością. Zatem nie można udowodnić, że taka gęstość jest faktycznie największa przez jak najciaśniejsze układanie kolejnych warstw kul, istnieje bowiem wiele różnych możliwości.

Choć doświadczenie właścicieli straganów z warzywami robi duże wrażenie – a układ regularny centrowany ściennie z pewnością można było zobaczyć na targowiskach już w predynastycznym Egipcie – to o niczym to jeszcze nie przesądza. Prawdę mówiąc, fakt, że metoda używana przez handlarzy warzywami jest dobrą odpowiedzią, to do pewnego stopnia dzieło przypadku. Właścicielom warzywniaków nie chodzi wcale o jak najciaśniejsze upakowanie pomarańczy w określonej przestrzeni – tak postawiony problem ma w zasadzie nieskończenie wiele możliwych rozwiązań. Im zależy na tym, żeby ułożyć pomarańcze w sposób stabilny na płaskiej powierzchni w świecie, w którym grawitacja działa w dół. Zupełnie naturalne jest więc, że zaczynają od ułożenia jednej warstwy, a potem układają na niej kolejną i tak dalej. Jeśli wkładają pomarańcze do prostokątnego pudła, to pierwsza warstwa będzie zapewne siecią kwadratową. Jeżeli natomiast układają je na powierzchni nieograniczonej z boków, to tak samo naturalny będzie układ kwadratowy, jak trójkątny. Tak się akurat składa, że oba prowadzą do uzyskania układu regularnego centrowanego ściennie – w przypadku warstw trójkątnych tak się stanie, jeśli kolejne będą układane w odpowiedni sposób. Sieć kwadratowa wydaje się w istocie kiepskim wyborem, ponieważ nie jest to najciaśniejszy możliwy sposób upakowania jednej warstwy. Jednak za sprawą przypadku, a nie przemyślanej decyzji, okazuje się, że nie ma to znaczenia.

Fizyków nie interesują pomarańcze. Bardziej zależy im na upakowaniu atomów. Kryształ jest regularnym, powtarzającym się okresowo układem atomów. Hipoteza Keplera wyjaśnia, że

okresowość wynika w naturalny sposób z dążenia do jak najciaśniejszego upakowania atomów. Z punktu widzenia większości fizyków istnienie kryształów jest wystarczającym dowodem na to, że hipoteza ta jest prawdziwa. Zauważyliśmy jednak przed chwilą, że istnieje nieskończenie wiele sposobów upakowania kul równie gęsto jak w przypadku układu regularnego centrowanego ściennie i układu heksagonalnego, ale żaden z tych alternatywnych układów nie charakteryzuje się okresową powtarzalnością. Dlaczego więc przyroda wykorzystuje w kryształach wzory okresowe? Być może odpowiedzią jest fakt, że kule nie oddają wiernie własności prawdziwych atomów.

Pomarańcze nie interesują też matematyków. Tak jak Kepler, wolą oni rozważać doskonałe, jednakowe kule. Nie przekonuje ich jednak argumentacja fizyków. Jeśli modelowanie atomów za pomocą kul jest niewłaściwe, to istnienie kryształów przestaje być dowodem przemawiającym za poprawnością hipotezy Keplera. Jedno wyklucza drugie. Nawet gdybyśmy próbowali dowodzić, że hipoteza ta w pewnym stopniu wyjaśnia sieci krystaliczne, a sieć krystaliczna do pewnego stopnia potwierdza poprawność hipotezy… to wciąż pozostanie pewna luka. Matematycy pragną uzyskać dowód.

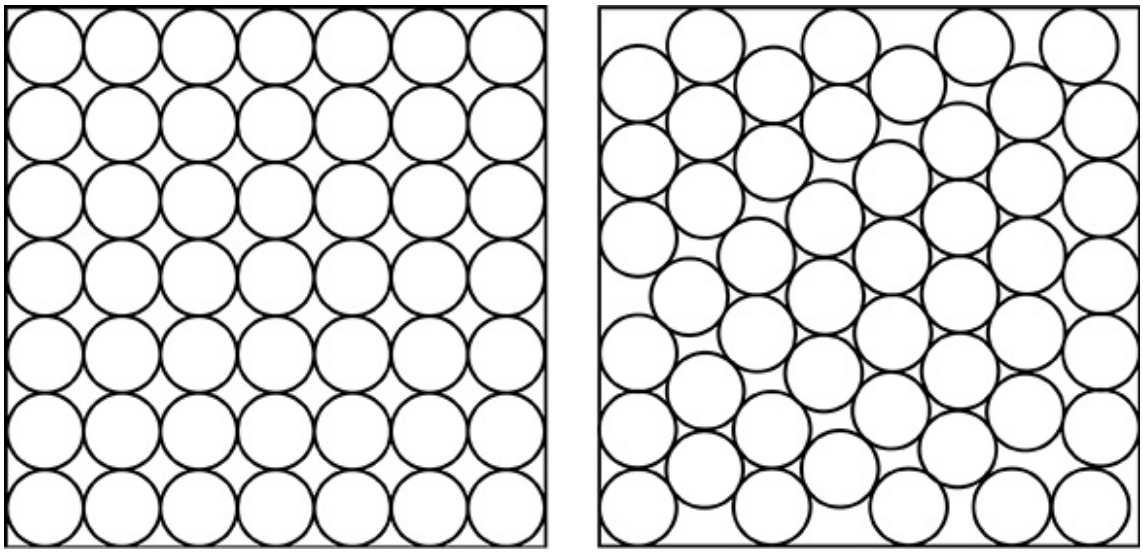
Kepler nie nazwał swojego spostrzeżenia hipotezą – zapisał ją jedynie w książce. Nie mamy najmniejszego pojęcia, czy faktycznie nadawał jej aż tak ogólne znaczenie. Czy chciał stwierdzić, że układ regularny centrowany ściennie jest „najciaśniejszym upakowaniem w trzech wymiarach” pośród wszystkich możliwych sposobów upakowania kul? A może miał na myśli jedynie to, że jest to najciaśniejsze upakowanie w gronie trzech rozważanych przez niego układów? Nie możemy się cofnąć w czasie i go zapytać. Bez względu na to, jak było naprawdę, dla matematyków i fizyków ważna jest interpretacja zakładająca takie ogólne, ambitne rozumienie. Takie, które wymaga rozważenia wszystkich możliwych sposobów upakowania nieskończonej liczby kul w nieskończonej przestrzeni – i udowodnienia, że żaden z tych sposobów nie prowadzi do uzyskania większej gęstości niż dla układu regularnego centrowanego ściennie.

Nietrudno popełnić błąd polegający na niedocenieniu hipotezy Keplera. To przecież zupełnie oczywiste, że uzyskanie największego upakowania wymaga układania kul w taki sposób, żeby każda kolejna dotykała jak największej liczby sąsiadek, prawda? W sposób nieunikniony prowadzi to do układu opisanego przez Keplera. To prawda, ale jedynie pod warunkiem że będziemy dodawali kolejne kule w odpowiedniej kolejności i umieszczali je we właściwych miejscach, gdy mamy do wyboru kilka możliwości. Nie ma jednak żadnej gwarancji, że nie może istnieć jakiś lepszy, bardziej przemyślany sposób od takiego dodawania kolejnych kul. Każdy, kto choć raz pakował walizki do bagażnika samochodowego, dobrze wie, że gdy po prostu wkłada się do niego kolejne pasujące pakunki, to mogą powstać luki, w których nie zmieści się już żadna torba. Wtedy pozostaje nam tylko jedno: trzeba dobrze wszystko przemyśleć i zacząć pakowanie od początku. Nie ulega wątpliwości, że pakowanie walizek jest takie skomplikowane częściowo dlatego, że bagaże mają różne rozmiary i kształty, ale trudno też nie zauważyć oczywistego wniosku: uzyskanie najciaśniejszego upakowania w jakimś małym obszarze może wywołać efekt domina prowadzący do tego, że nie uda nam się uzyskać takiego najlepszego upakowania w większej przestrzeni.

Kepler rozważał bardzo szczególne układy. Można sobie wyobrazić, że jakiś zupełnie inny układ mógłby pozwolić na jeszcze ciaśniejsze upakowanie takich samych kul. Może lepsze efekty przyniosłoby użycie nierównych warstw z wystającymi kulami? A może w ogóle należałoby zrezygnować z rozważania „warstw”? Poza tym, nawet jeśli mamy całkowitą pewność, że dana idea jest poprawna, i tak musimy to jeszcze udowodnić.

Nie przekonałem was? Wciąż uważacie, że to oczywiste? Tak oczywiste, że *nie wymaga* dowodu? Pozwólcie, że spróbuję zachwiać waszą wiarą we własną intuicję w kwestii upakowania kul. Rozważmy znacznie prostsze pytanie odnoszące się do kół na płaszczyźnie. Załóżmy, że dostaliście 49 jednakowych kół o średnicy wynoszącej 1 jednostkę. Jaką długość boku będzie miał *najmniejszy* kwadrat, w którym można je umieścić w taki sposób, żeby na siebie nie nachodziły? Na rycinie 21 (po lewej) pokazano rozwiązanie, które

od razu samo się narzuca: należy je upakować tak jak butelki mleka w skrzynce. Bok takiej skrzynki wynosi dokładnie 7 jednostek. Aby się upewnić, że jest to najlepszy układ, wystarczy zauważyć, że każde kółko tkwi nieruchomo na swoim miejscu, przytrzymywane pewnie przez sąsiednie koła, i w żaden sposób nie da się tu znaleźć dodatkowego miejsca. Na rycinie 21 (po prawej) pokazano, że jest to błędna odpowiedź. Po upakowaniu kółek w sposób nieregularny, tak jak przedstawiono na diagramie, zmieszczą się one w kwadracie o boku wynoszącym niecałe 6,98 jednostki[43]. Zatem nasz dowód również był błędny. Uzyskanie sztywnego, nieruchomego upakowania nie oznacza wcale, że nie można kółek upakować jeszcze lepiej.



Ryc. 21. Po lewej: 49 kółek w kwadracie 7 × 7. Po prawej: Jak zmieścić 49 kółek w nieco mniejszym kwadracie.

Prawdę mówiąc, nietrudno zauważyć, że rozumowanie prowadzące do wniosku, iż rozwiązaniem jest kwadrat 7 × 7, nie może być poprawne. Wystarczy rozważyć większe kwadraty. Jeśli użyjemy sieci kwadratowej, to n^2 kół o średnicy 1 upakujemy w kwadracie o boku n. W żaden sposób nie można zwiększyć gęstości przez przesuwanie kół w ciągły sposób, ponieważ takie upakowanie jest sztywne.

Jednak dla odpowiednio dużych n musi istnieć jakieś upakowanie dające większą gęstość, ponieważ sieć trójkątna jest pod tym względem lepsza od kwadratowej. Jeśli weźmiemy naprawdę duży kwadrat i upchniemy w nim jak najwięcej kół ułożonych w sieć trójkątną, to przewaga takiego sposobu układania nad siecią kwadratową doprowadzi w końcu do uzyskania większej gęstości mimo występowania „efektu brzegowego", czyli powstawania niemożliwych do uniknięcia luk przy bokach kwadratu. Długość całego brzegu wynosi $4n$ i może stać się dowolnie mała w porównaniu z n^2. Tak się akurat składa, że wartość, od której sieć trójkątna prowadzi do lepszego upakowania, to $n = 7$. Nie jest to wcale oczywiste i wyznaczenie tej wartości wymaga przeprowadzenia wielu szczegółowych obliczeń, ale przy jakiejś wartości n sieć trójkątna musiała uzyskać przewagę. Sama sztywność upakowania nie wystarcza.

Tak naprawdę istnieją dwie wersje hipotezy Keplera. Jedna z nich dotyczy tylko takich sieci, w których środki kul tworzą wzór powtarzający się okresowo bez końca w trzech niezależnych kierunkach niczym w jakiejś trójwymiarowej tapecie. Nawet przy takim założeniu udowodnienie hipotezy jest trudne, ponieważ w przestrzeni może istnieć wiele różnych sieci. Specjaliści od krystalografii wyróżniają 14 ich typów, pogrupowanych ze względu na posiadane symetrie. Niektóre z tych rodzajów sieci opisuje się za pomocą parametrów liczbowych, które mogą przyjmować nieskończenie wiele różnych wartości. To jednak nic w porównaniu z trudnościami, na jakie natrafimy, gdy rozważymy drugą wersję hipotezy, która dopuszcza wszystkie możliwe upakowania. Każda kula może unosić się w przestrzeni – nie działa tu grawitacja – i nic jej nie zmusza, by dołączyła do jakiejś warstwy czy innego układu symetrycznego.

Gdy jakiś problem staje się zbyt trudny, matematycy odkładają go na bok i rozważają jego prostsze wersje. Rozważania Keplera na temat płaskich warstw kul sugerują, że można by na początek zająć się układem kół na płaszczyźnie. Postawmy więc sobie takie zadanie: mając nieograniczony zapas jednakowych kół, spróbujmy je

upakować możliwie najściślej. Wówczas gęstość jest proporcjonalna do *pola* zajmowanego przez koła. W 1773 roku Joseph-Luis Lagrange udowodnił, że najgęstszym upakowaniem sieciowym kół na płaszczyźnie jest sieć trójkątna, której gęstość wynosi $\pi/\sqrt{12} = 0,9069$. W 1831 roku Gauss recenzował książkę Ludwiga Seebera, który uogólnił niektóre z twierdzeń Gaussa z zakresu teorii liczb na równania trzech zmiennych. Gauss zauważył, że wyniki Seebera dowodzą, iż układ regularny centrowany ściennie i układ heksagonalny stanowią najgęstsze upakowanie sieciowe w przestrzeni trójwymiarowej. Obecnie bardzo dużo już wiemy na temat upakowań sieciowych w przestrzeniach o większej liczbie wymiarów – 4, 5, 6 i tak dalej. Szczególnie dobrze rozumiemy przypadek przestrzeni 24-wymiarowej. (W matematyce tak to już bywa). Choć wydaje się, iż takie rozważania pozbawione są związku z rzeczywistością, to okazuje się jednak, że mają one duże znaczenie dla rozwoju teorii informacji i szyfrowania komputerowego.

Upakowania niesieciowe są czymś zupełnie innym. Jest ich nieskończenie wiele i nie mają regularnej struktury, wygodnej do analizy. Dlaczego więc nie mielibyśmy pójść na całość i rozważać upakowania przypadkowe? W swojej książce *Vegetable Staticks* (*Statyka warzyw*) z 1727 roku Stephen Hales opisał doświadczenia, w których „ściskał kilka paczek świeżego groszku w jednym naczyniu" i odkrył, że po takim ściśnięciu groszki zmieniają się w „piękne dwunastościany foremne". Ponieważ dwunastościany foremne nie mogą dokładnie wypełnić przestrzeni, Hales zapewne zaobserwował dwunastościany rombowe, które, jak już powiedzieliśmy, pojawiają się przy okazji upakowania regularnego centrowanego ściennie. G. David Scott umieścił natomiast w pojemniku dużą liczbę kulek łożyskowych, porządnie nim potrząsnął i stwierdził, że największa gęstość, jaką udało mu się uzyskać, wynosi 0,6366. W 2008 roku Chaoming Song, Ping Wang i Hernán Makse wyprowadzili tę wartość analitycznie[44]. Ich rezultat nie świadczy jednak o tym, że Kepler miał rację – choćby tylko dlatego, że wynikałoby z niego, iż układ regularny centrowany ściennie o gęstości 0,74 nie może istnieć. Najprostszym sposobem

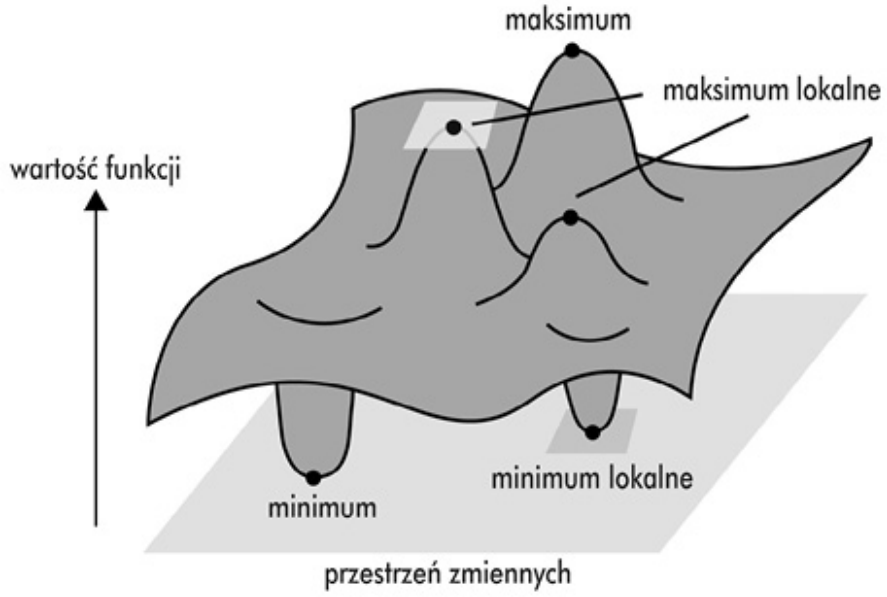
wyjaśnienia tej rozbieżności jest zwrócenie uwagi na fakt, że uczeni ci w swoich badaniach pominęli niezwykle rzadkie wyjątki. Tego rodzaju wyjątkami są właśnie układ regularny centrowany ściennie, układ heksagonalny i wszystkie przypadkowe układy warstw trójkątnych. Z tego samego powodu nie można odrzucić możliwości, że istnieją jakieś inne układy o jeszcze większej gęstości. Nie może to być żadna sieć, ale struktury sieciowej nigdy nie uzyskamy w sposób losowy, ponieważ prawdopodobieństwo jej wystąpienia jest praktycznie zerowe. Zatem badanie losowych upakowań, choć przydatne w rozważaniach wielu kwestii z dziedziny fizyki, nie powie nam zbyt wiele na temat hipotezy Keplera.

Pierwszy prawdziwy przełom nastąpił w 1892 roku, gdy Axel Thue wygłosił wykład na Skandynawskim Kongresie Nauk Przyrodniczych. W swoim wystąpieniu nakreślił szkic dowodu pokazującego, że żadne upakowanie kół na płaszczyźnie nie może być gęstsze od sieci trójkątnej. Jego wykład ukazał się drukiem, ale przedstawione w nim szczegóły są zbyt mgliste, by można było odtworzyć dowód, o którym mówił. W 1910 roku przedstawił za to nowy dowód, który wydawał się przekonujący – brakowało w nim tylko kilku szczegółowych kwestii, Thue jednak po prostu założył, że można je wyjaśnić. Zamiast zapełniać te luki, László Fejes Tóth uzyskał w 1940 roku zupełnie nowy dowód, przeprowadzony innymi metodami. Niedługo potem również Beniamino Segre i Kurt Mahler znaleźli alternatywne dowody. W roku 2010 Hai-Chau Chang i Lih-Chung Wang zamieścili w Internecie prostszy dowód[45].

Problem znalezienia największej gęstości upakowania kół lub kul w określonych warunkach należy do ogólniejszej klasy zagadnień matematycznych znanych jako problemy optymalizacji. Chodzi w nich o znalezienie wartości maksymalnej lub minimalnej określonej funkcji, czyli reguły matematycznej na obliczenie jakiejś wielkości zależnej od ustalonego zbioru zmiennych. Taką regułę zapisuje się bardzo często w postaci wzoru, ale nie jest to konieczne. W ten sposób można na przykład sformułować nasze zadanie z 49 kółkami, które należy umieścić wewnątrz kwadratu. Zmiennymi są w tym

przypadku współrzędne środków 49 kół, a ponieważ opisanie każdego koła wymaga podania dwóch współrzędnych, w sumie mamy 98 zmiennych. Wartością funkcji będzie rozmiar kwadratu o bokach równoległych do osi układu współrzędnych, w którym można zmieścić wszystkie kółka bez nakładania ich na siebie. W tej sytuacji rozwiązanie naszego zadania jest równoważne znalezieniu najmniejszej wartości, jaką tak zdefiniowana funkcja przyjmuje dla zmiennych z szerokiego zakresu wszystkich możliwych upakowań.

Funkcję można sobie wyobrazić jako wielowymiarowy krajobraz. Każdy punkt tego krajobrazu odpowiada określonym wartościom zmiennych, a jego wysokość jest odzwierciedleniem wartości funkcji. Wartość maksymalna funkcji jest zatem wysokością najwyższego szczytu, a minimalna – głębokością najgłębszej doliny. Zadania z dziedziny optymalizacji można w zasadzie rozwiązać za pomocą rachunku różniczkowego i całkowego – na wierzchołku szczytu i na dnie doliny funkcja musi być pozioma (zob. ryc. 22), a rachunek różniczkowy i całkowy pozwala wyrazić to wymaganie w postaci równania. Uzyskanie tą metodą rozwiązania naszego zadania z 49 kółkami wymagałoby rozwiązania układu 98 równań z 98 zmiennymi.



Ryc. 22. Góry i doliny funkcji

Z problemami z dziedziny optymalizacji wiąże się ten kłopot, że tego typu równania mają często olbrzymią liczbę rozwiązań. W krajobrazie może występować wiele lokalnych szczytów, ale tylko jeden z nich będzie największy. Wyobraźmy sobie Himalaje – w tej okolicy nie ma nic *prócz* szczytów, a jednak rekord wysokości należy do Mount Everestu. Różne metody ich znajdowania – z których najbardziej oczywistą jest: „zmierzaj pod górę tak długo, jak się da" – często grzęzną na jakimś lokalnym szczycie. Inny szkopuł polega na tym, że wraz ze wzrostem liczby zmiennych rośnie również liczba lokalnych szczytów. Niemniej metoda ta czasami działa. Nawet częściowe wyniki mogą być użyteczne – jeśli uda się znaleźć jakiś lokalny szczyt, to będzie przynajmniej wiadomo, że maksimum musi mieć co najmniej taką wartość. Właśnie w ten sposób znaleziono lepszy układ kółek w kwadracie.

W wypadku układów sieciowych analizowana funkcja zależy tylko od skończonej liczby zmiennych, kierunków i długości, wzdłuż których sieć się powtarza. Gdy rozważamy upakowania, które nie są sieciami, funkcja zależy od nieskończenie wielu zmiennych – są nimi środki wszystkich kół lub kul. W takich wypadkach nie można bezpośrednio użyć rachunku różniczkowego i całkowego lub innych technik optymalizacji. W swoim dowodzie Tóth wykorzystał sprytny pomysł pozwalający przekształcić niesieciowy problem upakowania kół w problem optymalizacyjny *skończonego* zbioru zmiennych. Później, w 1953 roku, Tóth uświadomił sobie, że tę samą sztuczkę można by w zasadzie zastosować do hipotezy Keplera. Niestety, uzyskana w ten sposób funkcja zależy od około 150 zmiennych, a to o wiele za dużo, by takie równania można było rozwiązać ręcznie. Tóth przewidział jednak proroczo wyjście z tej sytuacji: „Biorąc pod uwagę szybki rozwój komputerów, można sobie wyobrazić, że wartość minimalną uda się wyznaczyć z bardzo dużą dokładnością".

W tamtym okresie technika komputerowa dopiero raczkowała i nie istniało żadne urządzenie, które mogłoby sobie poradzić z takim zadaniem. Dalsze postępy w pracach nad hipotezą Keplera uzyskano więc dzięki zastosowaniu innego podejścia. Różnym matematykom udało się wyznaczyć górne ograniczenia na gęstość upakowania kul. Na przykład w 1958 roku Rogers udowodnił, że gęstość ta wynosi

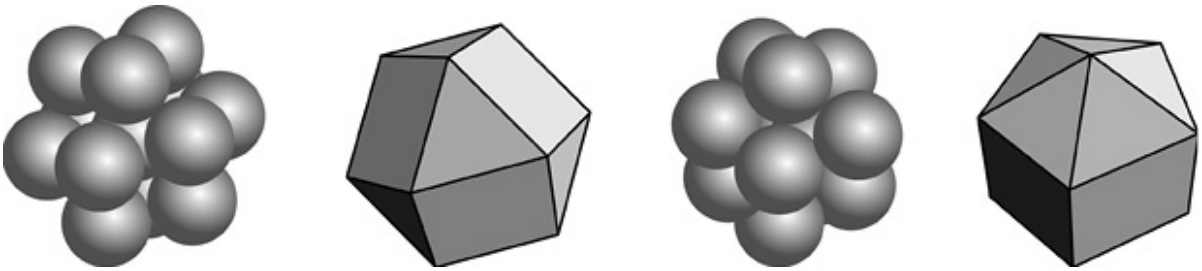
najwyżej 0,7797 – tu nie ma żadnych wyjątków: ograniczenie dotyczy wszystkich upakowań kul. W 1986 roku J.H. Lindsey przesunął to ograniczenie do wartości 0,77844, a w 1988 roku Douglasowi Muderowi[46] udało się uszczknąć jeszcze trochę i otrzymał ograniczenie wynoszące 0,77836. Wyniki te pokazują, że nie można uzyskać *dużo* większej gęstości niż w wypadku układu regularnego centrowanego ściennie, dla którego gęstość wynosi 0,7405. Wciąż jednak istniała pewna luka, a szanse na jej wypełnienie były niewielkie.

W 1990 roku amerykański matematyk Wu-Yi Hsiang oznajmił, że znalazł dowód hipotezy Keplera. Gdy jednak opublikował szczegóły swojej pracy, szybko zaczęły się pojawiać wątpliwości. Tóth zamieścił w czasopiśmie „Mathematical Reviews" recenzję, w której napisał: „Gdyby ktoś mnie zapytał [czy w tym artykule znajduje się] dowód hipotezy Keplera, odpowiedziałbym: nie. Mam nadzieję, że Hsiang uzupełni brakujące szczegóły, ale coś mi mówi, że lwia część tej pracy nie została jeszcze wykonana".

Thomas Hales, który zajmował się tą hipotezą przez wiele lat, również wątpił, czy metodę Hsianga można poprawić. Zamiast tego postanowił poświęcić czas na poważną analizę podejścia Tótha. Dorosło już nowe pokolenie matematyków, dla których sięgnięcie po komputer było bardziej naturalne od zajrzenia do tablic logarytmicznych. W 1996 roku Hales nakreślił szkic strategii przeprowadzenia dowodu opartego na pomyśle Tótha. Wymagało to znalezienia wszystkich możliwych sposobów ułożenia kilku kul w bezpośrednim sąsiedztwie jednej wybranej. Układ kul można opisać, podając położenie ich środków – w wypadku kul jednostkowych ich środki muszą być odległe od siebie przynajmniej o 2 jednostki. Przyjmijmy, że dwie kule są *sąsiednie*, jeśli ich środki są oddalone od siebie najwyżej na odległość 2,51 jednostki. Przyjęcie odpowiedniej wartości jest kwestią wyczucia: jeśli weźmiemy zbyt małą wartość, to nie będzie wystarczająco dużo miejsca na przestawienie sąsiednich kul i zwiększenie gęstości, a jeśli przyjmiemy zbyt dużą wartość, to liczba możliwych sposobów

przestawienia kul będzie ogromna. Hales ustalił, że wartość 2,51 jest dobrym kompromisem. Możemy teraz przedstawić układ sąsiednich kul, tworząc w przestrzeni nieskończoną sieć. Jej węzły znajdują się w środkach kul, i są połączone krawędziami tylko wtedy, gdy odpowiadające im kule są sąsiadkami. Sieć taka jest swego rodzaju szkieletem upakowania, zawierającym niezwykle ważne informacje na temat sąsiedztwa każdej kuli.

Dla każdej wybranej kuli możemy rozważać jej bezpośrednie sąsiadki w sieci i brać pod uwagę tylko łączące je krawędzie, pomijając wybraną na początku kulę. W efekcie uzyskamy pewien rodzaj klatki otaczającej wierzchołek znajdujący się w środku wybranej kuli. Na rycinie 23 (po lewej) pokazano sąsiadki jednej z kul w układzie regularnym centrowanym ściennie i odpowiadającą im klatkę. Na tej samej rycinie po prawej stronie zamieszczono analogiczne rysunki dla szczególnego układu kul tworzącego graniastosłup pięciokątny, który okazał się szczególnie ważny w omawianym dowodzie. W tym wypadku mamy dwa pięciokąty ułożone równolegle do „równika” kuli środkowej i po jednej kuli na każdym biegunie.



Ryc. 23. Od lewej: Sąsiedztwo kuli w układzie regularnym centrowanym ściennie; klatka utworzona przez jej sąsiadki; sąsiedztwo kuli o układzie graniastosłupa pięciokątnego; klatka utworzona przez jej sąsiadki.

Klatki tworzą bryły o płaskich ścianach i geometria tych brył decyduje o gęstości upakowania w pobliżu kuli znajdującej się w ich środku[47]. Kluczowy pomysł polega na przypisaniu każdej takiej

klatce odpowiedniej liczby, którą nazywamy jej wynikiem. Liczbę tę można interpretować jako oszacowanie gęstości upakowania sąsiadek danej kuli. Wynik nie jest wartością gęstości, ale wielkością, która jest bardziej przewidywalna i łatwiejsza do obliczenia. W szczególności wynik klatki można wyznaczyć, sumując po prostu wyniki związane z jej ścianami, co nie musi już być prawdą dla gęstości. W ogólnym przypadku wiele różnych definicji wyniku spełnia ten warunek, ale wszystkie zgadzają się w jednym: dla układu regularnego centrowanego ściennie i układu heksagonalnego wynik wynosi zawsze 8 „punktów", bez względu na to, jaką definicję wybierzemy. Punkt ma ściśle określoną wartość:

$$4\,\mathrm{arctg}\,\frac{\sqrt{2}}{5}-\frac{\pi}{3}=0,0553736$$

Zatem 8 punktów to w istocie 0,4429888. Ta dziwna liczba wynika ze szczególnej geometrii układu regularnego centrowanego ściennie. Hales zauważył pewien kluczowy fakt, który wiąże hipotezę Keplera z tą liczbą: jeśli każda klatka ma wynik wynoszący 8 punktów lub mniej, to hipoteza Keplera jest prawdziwa. Matematycy skupili się więc teraz na klatkach i wynikach.

Klatki można pogrupować ze względu na ich topologię, czyli ze względu liczbę ścian o danej liczbie stron i na sposób połączenia tych ścian ze sobą. Jednak klatki zaliczane do jednej topologii mogą mieć krawędzie o wielu różnych długościach. Długość krawędzi wpływa na wynik, ale jedna topologia grupuje ze sobą wiele różnych klatek i dzięki temu można się nimi wszystkimi zająć w taki sam ogólny sposób. W swoim ostatecznym dowodzie Hales rozważył około 5000 różnych rodzajów klatek, ale w głównych obliczeniach skupił się na kilkuset. W 1992 roku rozpoczął realizację pięcioetapowego programu:

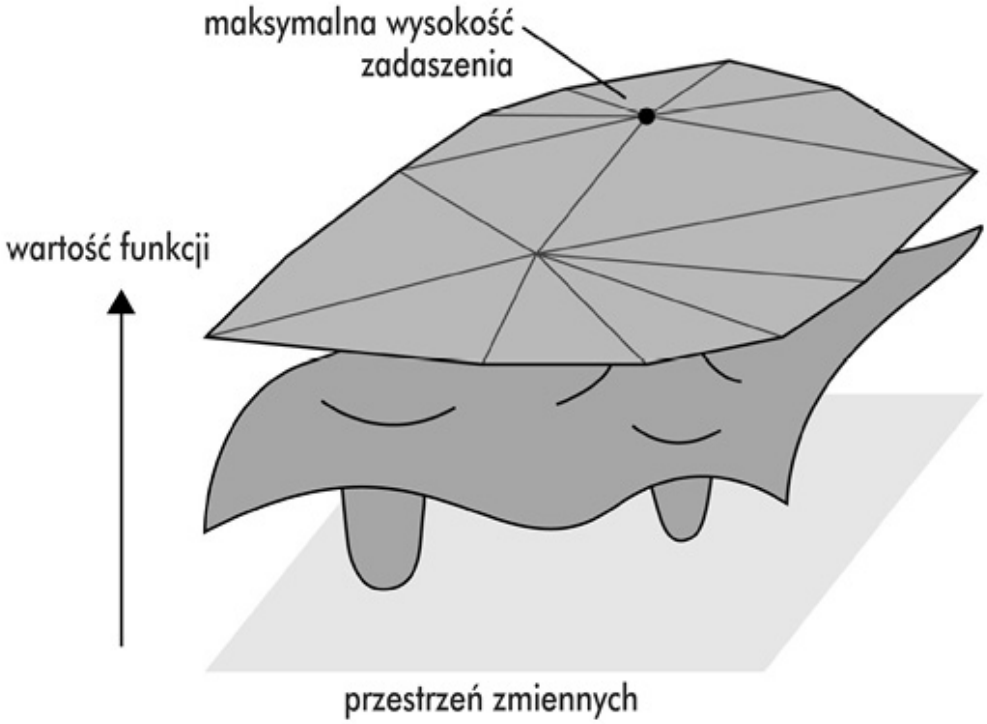
1. Udowodnienie istnienia spodziewanej wartości dla przypadków, gdy wszystkie ściany klatki są trójkątami.

2. Udowodnienie, że upakowanie regularne centrowane ściennie i upakowanie heksagonalne mają wyższą wartość wyniku niż jakakolwiek inna klatka o takiej samej topologii.
3. Uporanie się z sytuacjami, w których wszystkie ściany klatek są trójkątami i czworokątami, z wyjątkiem trudniejszego przypadku graniastosłupa pięciokątnego.
4. Uporanie się z dowolnymi klatkami, których ściany mają więcej niż cztery krawędzie.
5. Rozwiązanie ostatniego przypadku, który pozostał, czyli klatki w kształcie graniastosłupa pięciokątnego.

Pierwszy punkt został wykonany w 1994, a drugi w 1995 roku. Podczas realizacji tego programu Hales zmodyfikował definicję klatki, tak by uprościć cały wywód (sam używał określenia „gwiazda dekompozycyjna"). Nowa definicja nie zmienia jednak obydwu klatek pokazanych na rycinie 23 i nie wpłynęła też w żaden istotny sposób na te części dowodu, które zostały już zrealizowane. Dzięki zastosowaniu tego nowego pojęcia udało mu się do 1998 roku zakończyć realizację wszystkich pięciu kroków programu. Student Halesa Samuel Ferguson poradził sobie z punktem piątym, czyli ze skomplikowanym przypadkiem graniastosłupa pięciokątnego.

Badania te wymagały wykorzystania komputerów na wszystkich etapach. Cała sztuka polega na wybraniu dla każdej sieci lokalnej takiej definicji wyniku, żeby obliczenia były względnie proste. Z geometrycznego punktu widzenia zastąpienie gęstości wartością punktową odpowiada zbudowaniu swego rodzaju zadaszenia nad gładkim krajobrazem, w którym szukamy szczytu. Owo zadaszenie składa się z wielu figur płaskich (zob. ryc. 24). Analiza tego rodzaju kształtów jest prostsza niż gładkich powierzchni, ponieważ wartości maksymalne muszą być w rogach, a te można znaleźć, rozwiązując dużo prostsze równania. Istnieją skuteczne metody przeprowadzania takich obliczeń wywodzące się z dziedziny programowania liniowego. Jeśli zadaszenie skonstruowano sprytnie w taki sposób, że jego szczyt zbiega się ze szczytem gładkiej powierzchni pod spodem, to

takie prostsze obliczenia pozwalają znaleźć szczyt gładkiej powierzchni.



Ryc. 24. Zadaszenie nad funkcją

Niestety takie podejście ma swoją cenę: trzeba rozwiązać około 100 000 problemów z zakresu programowania liniowego. Wymaga to przeprowadzenia długich obliczeń, które nie wykraczają jednak poza możliwości współczesnych komputerów. Gdy Hales i Ferguson przygotowali swoją pracę do publikacji, liczyła ona około 250 stron równań matematycznych plus 3 gigabajty plików komputerowych.

W 1999 roku Hales przesłał dowód do czasopisma „Annals of Mathematics", którego redakcja powołała specjalnie do oceny tego artykułu panel 12 ekspertów. W 2003 roku członkowie panelu ogłosili, że mają „99 procent pewności", iż dowód jest poprawny. Pozostała jeszcze niepewność związana z obliczeniami komputerowymi. Recenzenci powtórzyli wiele z nich i sprawdzili, że

metoda przeprowadzenia dowodu jest uporządkowana, a programy są dobrze napisane, ale nie mogli zweryfikować niektórych aspektów dowodu. Z pewnym opóźnieniem artykuł ukazał się w końcu w „Annals of Mathematics". Hales przyznał, że takiego podejścia do przeprowadzenia dowodu zapewne nigdy nie uda się zweryfikować w stu procentach, i w 2003 roku ogłosił, że rozpoczyna prace nad przekształceniem dowodu do takiej postaci, którą można będzie sprawdzić komputerowo z wykorzystaniem standardowego oprogramowania do automatycznej weryfikacji dowodów.

Mogłoby się wydawać, że proponując to, Hales wpakował się z deszczu pod rynnę, ale taki plan jest całkowicie sensowny. Dowody publikowane w czasopismach mają przekonać ludzi. Jak powiedzieliśmy w rozdziale 1, tego rodzaju dowód jest opowieścią. Komputery są kiepskie w snuciu opowieści, ale za to są niedoścignione w tym, w czym my sami jesteśmy beznadziejni: w przeprowadzaniu długich, żmudnych obliczeń bez popełniania błędów. Komputery świetnie nadają się do sprawdzania dowodów przygotowanych zgodnie z ich podręcznikową definicją, czyli będących ciągiem logicznych kroków wynikających z siebie nawzajem.

Informatycy zdołali wykorzystać tę umiejętność komputerów. Jeśli chcemy sprawdzić dowód, wystarczy zweryfikować na komputerze każdy krok logiczny. Powinno to być całkiem łatwe, ale niestety autorzy dowodów publikowanych w czasopismach naukowych nie zapisują ich w taki sposób. Pomijają wszystkie rutynowe lub oczywiste kroki. Często pojawiają się tam tradycyjne sformułowania świadczące o tym, że coś pominięto: „Łatwo można sprawdzić, że…", „Wykorzystując metody Cheesbergera i Frytka, zmodyfikowane tak, by uwzględniały pojedyncze osobliwości, widzimy, że…", „Po krótkich obliczeniach mamy…". Komputery nie potrafią (jeszcze) radzić sobie w takich sytuacjach. Ludzie mogą jednak przepisać dowody na nowo, wypełniając tego typu luki, a wtedy komputery umieją już sprawdzić każdy krok.

Powód, dla którego nie oznacza to wpadnięcia z deszczu pod rynnę, jest dość prosty: oprogramowanie sprawdzające dowody trzeba przetestować *tylko raz*. Jest to program ogólny, który można

stosować do wszystkich dowodów zapisanych w odpowiedniej postaci. Wszystkie obawy na temat dowodów komputerowych skupiają się na tym jednym programie. Wystarczy się upewnić, że jest on poprawny, a wtedy będziemy mogli sprawdzać poprawność wielu innych dowodów. Można nawet zautomatyzować również ten proces, pisząc program do badania dowodów w języku, który daje się sprawdzić znacznie prostszym programem do kontroli poprawności oprogramowania.

W ostatnich latach zweryfikowano w ten sposób dowody wielu ważnych twierdzeń matematycznych. Często wymagało to przedstawienia ich w postaci, która jest bardziej odpowiednia do przetwarzania na komputerze. Jednym z najnowszych osiągnięć tego rodzaju jest sprawdzenie dowodu twierdzenia o krzywej Jordana: każda nieprzecinająca się krzywa zamknięta na płaszczyźnie dzieli ją na dwa różne obszary spójne. Może się wydawać, że to oczywiste, ale twórcy topologii długo nie mogli znaleźć ścisłego dowodu tego twierdzenia. W końcu w 1887 roku Camille Jordan przedstawił osiemdziesięciostronicowy dowód, ale później zarzucono mu, że poczynił w nim nieuzasadnione założenia. Ostatecznie palma pierwszeństwa przypadła Oswaldowi Veblenowi, który w 1905 roku zaprezentował bardziej szczegółowy dowód, mówiąc: „Dowód [Jordana] […] jest niezadowalający w opinii wielu matematyków. Zakłada on poprawność twierdzenia, nie przedstawiając dowodu w najważniejszym przypadku szczególnym wielokąta prostego i wynikającej z tego argumentacji. Należy więc stwierdzić przynajmniej to, że nie podano w nim wszystkich szczegółów”. Matematycy przyjęli krytyczne słowa Veblena bez sprzeciwu, ale niedawno Hales przejrzał ponownie dowód Jordana i nie znalazł w nim „nic, co budziłoby sprzeciw”. Prawdę mówiąc, uwaga Veblena na temat wielokąta jest dziwna: twierdzenie to jest oczywiste w przypadku wielokąta, a poza tym dowód Jordana i tak nie opiera się na takiej wersji[48]. Z dowodami opowiadającymi pewną historię wiążą się specyficzne niebezpieczeństwa. Zawsze warto sprawdzić, czy popularna wersja danej opowieści jest taka sama jak oryginał.

W ramach rozgrzewki przed zmaganiami z hipotezą Keplera Hales przedstawił w 2007 roku sprawdzony komputerowo dowód formalny twierdzenia o krzywej Jordana. Wykorzystał do tego program zawierający 60 000 linii. Niedługo potem pewien zespół matematyków pokazał inny dowód formalny przeprowadzony z wykorzystaniem innego programu. Komputerowe wspomaganie procesu dowodzenia nie zapewnia całkowitej bezbłędności, ale gwarancji takiej nie daje również podejście tradycyjne. Szczerze mówiąc, w wielu artykułach naukowych z dziedziny matematyki znajduje się gdzieś zapewne jakiś błąd rachunkowy. Od czasu do czasu komuś udaje się wykryć takie pomyłki i najczęściej okazuje się, że nie mają one większego znaczenia. Poważne błędy znajduje się zazwyczaj dlatego, że wprowadzają niespójności i widać, iż coś się nie zgadza. Oto kolejna wada dowodów snujących opowieści: ceną, jaką musimy płacić za to, że dowody są zrozumiałe dla ludzi, jest to, że wątek może być czasami bardzo przekonujący, nawet jeśli prowadzi na manowce.

Hales nadał swojemu nowemu przedsięwzięciu nazwę *Project FlysPecK* – litery F, P oraz K są skrótem od ang. *Formal Proof of Kepler* (formalny dowód [hipotezy] Keplera)[49]. Początkowo szacował, że realizacja tego projektu zajmie około dwudziestu lat[45]. Po upływie dziewięciu lat od rozpoczęcia badań udało się już dokonać znacznych postępów. Niewykluczone, że prace uda się zakończyć wcześniej.

36 Johannes Kepler, *Noworoczny podarek albo O sześciokątnych płatkach śniegu*, przeł. Dorota Sutkowska, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2013 (przyp. tłum.).

37 Johannes Kepler, *Noworoczny podarek albo O sześciokątnych płatkach śniegu*, przeł. Dorota Sutkowska, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2013, s. 49 (przyp. tłum.).

38 *Ibid.*, s. 53 (przyp. tłum.).

39 *Ibid.*, s. 54 (przyp. tłum.).

40 Wiele fascynujących informacji na temat fizyki płatków śniegu można znaleźć (w języku angielskim) na stronie http://www.its.caltech.edu/~atomic/snowcrystals/.

41 Claude Ambrose Rogers, *The packing of equal spheres* (*Upakowanie jednakowych kul*), „Proceedings of the London Mathematical Society" 1958, tom 8, s. 609–620.

42 Ponieważ przestrzeń jest nieskończona, mamy nieskończenie wiele kul i zarówno przestrzeń, jak i kule zajmują nieskończoną całkowitą objętość. Nie możemy jednak przyjąć, że gęstość wynosi ∞/∞, ponieważ wyrażenie to nie ma dobrze określonej wartości liczbowej. Zamiast tego rozważamy stopniowo coraz większe obszary przestrzeni i za miarę upakowania przyjmujemy wartość graniczną wyniku z dzielenia objętości zajmowanej przez kule w danym obszarze przez objętość całego obszaru.

43 Zob.: http://hydra.nat.uni-magdeburg.de/packing/csq/csq49.html.

44 Chaoming Song, Ping Wang, Hernán Makse, *A phase diagram for jammed matter* (*Diagram fazowy materii w stanie zakorkowania*), „Nature" nr 453, 29 maja 2008, s. 629–632.

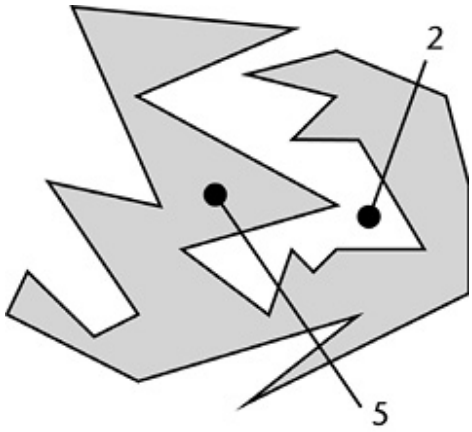
45 Hai-Chau Chang, Lih-Chung Wang, *A simple proof of Thue's theorem on circle packing* (*Prosty dowód twierdzenia Thue'a o upakowaniu kół*), arXiv:1009.4322v1, 2010.

46 J.H. Lindsey, *Sphere packing in* $\Re^3$ (*Upakowanie kul w* $\Re^3$), „Mathematika" 1986, tom 33, s. 137–147. Douglas J. Muder, *Putting the best face on Voronoi polyhedron* (*Najlepszy wygląd wielościanu Woronoja*), „Proceedings of the London Mathematical Society" 1988, tom 56, s. 329–348.

47 Hales posługiwał się kilkoma różnymi określenia na to, co my nazywamy tu klatką. Ostatecznie wybrał termin: „gwiazda dekompozycyjna". Nasz opis pomija niektóre kluczowe rozróżnienia, ale dzięki temu podstawowa idea jest bardziej zrozumiała.

48 Rozważmy obszar w kształcie wielokąta, taki jak na rycinie 55. Dla każdego punktu nienależącego do wielokąta możemy narysować wybiegającą z niego półprostą, która będzie wychodziła poza okrąg zawierający wielokąt bez

przechodzenia przez którykolwiek z wierzchołków wielokąta. (Istnieje skończona liczba wierzchołków, natomiast półprostych mamy do wyboru nieskończenie wiele). Taka półprosta przecina boki wielokąta skończoną liczbę razy i liczba ta będzie parzysta lub nieparzysta. Zdefiniujmy wnętrze wielokąta jako zbiór wszystkich punktów, dla których liczba ta jest nieparzysta, natomiast zewnętrzną część będą tworzyły punkty o parzystej liczbie przecięć boków wielokąta. Przy takich definicjach nietrudno udowodnić, że każdy z tych obszarów jest spójny i że są one od siebie oddzielone bokami wielokąta.



Ryc. 55. Dowód twierdzenia o krzywej Jordana dla wielokąta. W przypadku punktów znajdujących się na zacienionym obszarze (wnętrze) mamy nieparzystą liczbę przecięć krzywej, natomiast półproste poprowadzone z punktów na białym obszarze (zewnętrzna część) przecinają krzywą Jordana parzystą liczbę razy.

49 Nazwę projektu znaleziono, przeglądając angielski słownik w poszukiwaniu słowa zaczynającego się na literę „f” i zawierającego następnie „p” i „k”. Słowo *flyspeck* oznacza maleńką czarną plamkę, jaką pozostawiają po sobie muchy (przyp. tłum.).

6. Nowe rozwiązania starego problemu

Hipoteza Mordella

Zbliżamy się już powoli do wielkiego twierdzenia Fermata, pora więc, byśmy powrócili na grunt teorii liczb. Aby się odpowiednio przygotować, zajmijmy się najpierw mniej znanym, ale chyba ważniejszym problemem. W 2002 roku Andrew Granville i Thomas Tucker opisali go następująco[50]:

> W 1922 roku Mordell napisał jeden z najważniejszych artykułów w historii matematyki. […] Na samym jego końcu postawił pięć pytań, które dostarczyły motywacji do prowadzenia ważnych badań z zakresu równań diofantycznych w XX wieku. Odpowiedź na najważniejsze i najtrudniejsze z tych pytań poznaliśmy w 1983 roku dzięki odkryciom Faltingsa, które okazały się jednymi z najgłębszych i najpotężniejszych idei w historii matematyki.

Zacznijmy od rozszyfrowania padających tu nazwisk: Louis Mordell był brytyjskim teoretykiem liczb, który urodził się w Stanach Zjednoczonych w żydowskiej rodzinie o litewskich korzeniach, natomiast Gerd Faltings jest niemieckim matematykiem. Pytanie, o którym mówią Granville i Tucker, było znane jako hipoteza Mordella, ale z przytoczonego cytatu można się domyślić, że obecnie nie jest to już hipoteza, ale genialnie udowodnione twierdzenie Faltingsa.

Hipoteza Mordella należy do ważnej dziedziny teorii liczb, jaką są równania diofantyczne. Nazwano je tak na cześć Diofantosa z Aleksandrii, który około 250 roku n.e. napisał słynną książkę *Arithmetica* (*Arytmetyka*). Historycy sądzą, że *Arithmetica* składała się z 13 ksiąg, ale do naszych czasów przetrwało jedynie sześć z nich i wszystkie w późniejszych odpisach. Nie jest to dzieło poświęcone arytmetyce rozumianej jako dodawanie i mnożenie liczb. Jest to pierwszy w historii podręcznik algebry, w którym Diofantos zgromadził większość tego, co starożytni Grecy wiedzieli na temat rozwiązywania równań. Znajduje się w niej nawet podstawowa forma notacji algebraicznej, w której, jak sądzimy, używano jednej z odmian greckiej litery ς (sigma) na oznaczenie niewiadomej (nasze *x*), zapisu Δ^{Υ} na oznaczenie jej kwadratu (nasze x^2), a zapis K^{Υ} oznaczał jej sześcian (nasze x^3). Dodawanie zapisywano, umieszczając symbole obok siebie, odejmowanie miało własny specjalny symbol, odwrotność niewiadomej (nasze 1/*x*) zapisywano jako ς^{χ} i tak dalej. Wszystkie te symbole odtworzono na podstawie późniejszych odpisów i przekładów, niewykluczone więc, że zakradły się tu jakieś błędy.

Zgodnie z prawidłami matematyki starożytnych Greków rozwiązania równań omawiane w książce *Arithmetica* musiały być liczbami wymiernymi, czyli ułamkami utworzonymi z liczb całkowitych, takimi jak 22/7. Nierzadko wymagano wręcz, by *były* liczbami całkowitymi. Rozważano tylko liczby dodatnie – liczby ujemne wprowadzono dopiero kilka wieków później w Chinach i Indiach. Obecnie tego typu równania nazywamy równaniami diofantycznymi. W książce Diofantosa można znaleźć kilka niezwykle ważnych spostrzeżeń. W szczególności wydaje się, że Diofantos miał świadomość tego, iż każdą liczbę całkowitą można wyrazić jako sumę kwadratów czterech liczb całkowitych (z uwzględnieniem zera). Twierdzenie to udowodnił Lagrange w 1770 roku. W tym miejscu interesuje nas inne równanie, a mianowicie wzór na wszystkie trójki pitagorejskie, czyli takie trójki liczb, że suma kwadratów dwóch z nich jest kwadratem trzeciej. Ich nazwa wywodzi się z twierdzenia Pitagorasa – taka równość wiąże ze sobą boki trójkąta prostokątnego. Najbardziej znanym przykładem

jest słynny trójkąt 3–4–5: $3^2 + 4^2 = 5^2$. Inny przykład to: $5^2 + 12^2 = 13^2$. Istnieje nieskończenie wiele takich trójek pitagorejskich, a przepis na znalezienie ich wszystkich pojawia się w dwóch lematach (twierdzeniach pomocniczych) poprzedzających twierdzenia 29 i 30 z X księgi *Elementów* Euklidesa.

Metoda Euklidesa pozwala otrzymać nieskończenie wiele trójek pitagorejskich. Mordell znał kilka innych równań diofantycznych, dla których istnieje wzór dający nieskończenie wiele rozwiązań. Znał również inny rodzaj równania diofantycznego o nieskończenie wielu rozwiązaniach, których *nie* opisuje jednak żaden wzór. Są to tak zwane krzywe eliptyczne – to dość dziwaczna nazwa, wziąwszy pod uwagę fakt, że nie mają one praktycznie nic wspólnego z elipsami – i w ich przypadku mamy do czynienia z nieskończoną liczbą rozwiązań, ponieważ w celu otrzymania kolejnego rozwiązania wystarczy ze sobą połączyć dwa inne. Mordell udowodnił jedną z podstawowych własności tych równań: aby uzyskać w ten sposób wszystkie rozwiązania, wystarczy mieć na początku ich skończoną liczbę.

Wyjąwszy te dwa rodzaje równań, każde inne równanie diofantyczne, jakie Mordell potrafił sobie wyobrazić, pasowało do jednej z dwóch kategorii: albo było wiadomo, że ma skończoną liczbę rozwiązań (do tej grupy zaliczamy również równania, które w ogóle nie mają rozwiązań), albo nikt nie wiedział, czy liczba rozwiązań jest skończona, czy nieskończona. Spostrzeżenie to, samo w sobie, nie było niczym nowym, ale Mordell zauważył pewną prawidłowość, na którą nikt wcześniej nie zwrócił uwagi. Prawidłowość ta nie ma nic wspólnego z teorią liczb – jej źródłem jest topologia. Ważne jest mianowicie to, ile otworów ma dane równanie. Poza tym, aby to wszystko nabrało sensu, należało jeszcze przyjąć, że rozwiązania są liczbami zespolonymi, a nie wymiernymi czy całkowitymi. Takie postawienie sprawy w jakiś sposób wydawało się sprzeczne z duchem równań diofantycznych.

Warto w tym miejscu poświęcić trochę czasu na zagłębienie się w szczegóły. Będą one niezwykle pomocne w dalszej części naszej

opowieści. Mam nadzieję, że nie odstraszą was zamieszczone tu równania – służą one jedynie temu, byśmy mogli skupić uwagę na konkretnym przykładzie. Najważniejsza jest bowiem opowieść wynikająca z tych rozważań algebraicznych.

Trójki pitagorejskie są rozwiązaniami w dziedzinie liczb całkowitych równania Pitagorasa:

$$x^2 + y^2 = z^2$$

Po podzieleniu przez z^2 mamy:

$$(x/z)^2 + (y/z)^2 = 1$$

Jak dowiedzieliśmy się w rozdziale 3, wynika z tego, iż występująca w równaniu para liczb wymiernych (x/z, y/z) leży na okręgu jednostkowym na płaszczyźnie. Równanie pitagorejskie wywodzi się z geometrii, gdzie opisuje trójkąt prostokątny. Wyprowadzony przed chwilą wzór stanowi nieco inną interpretację geometryczną, która nie odnosi się do jednej konkretnej trójki pitagorejskiej, ale do wszystkich możliwych trójek. Rozwiązania równania Pitagorasa odpowiadają bezpośrednio i naturalnie wszystkim punktom wymiernym na okręgu jednostkowym. Przez punkt wymierny rozumiemy tu taki punkt, którego obie współrzędne są liczbami wymiernymi.

Z tego związku można się dowiedzieć wielu interesujących rzeczy. Odrobina trygonometrii lub parę przekształceń algebraicznych pozwala odkryć, że dla dowolnej liczby t punkt

$$\left(\frac{2t}{t^2+1}, \frac{t^2-1}{t^2+1} \right)$$

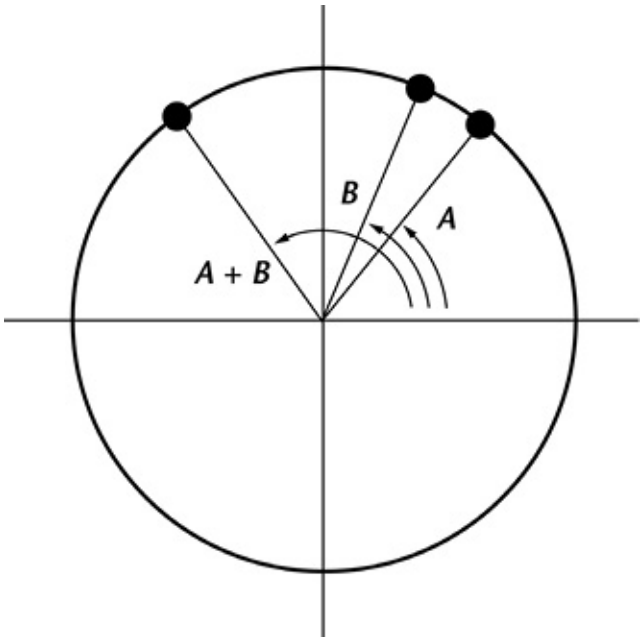
leży na okręgu jednostkowym. Co więcej, jeśli t jest liczbą wymierną, to również ten punkt będzie wymierny. W taki sposób możemy

uzyskać wszystkie punkty wymierne, zatem tym samym otrzymaliśmy wzór na wszystkie rozwiązania równania pitagorejskiego. Jest to wzór równoważny temu, jaki znajdziemy w dziele Euklidesa – ten sam przytacza również Diofantos. Jeśli przyjmiemy na przykład, że $t = 22/7$, to z naszego wzoru otrzymamy punkt

$$\left(\frac{308}{533}, \frac{435}{533}\right)$$

i sami możecie sprawdzić, że $308^2 + 435^2 = 533^2$. Dla naszych rozważań dokładna postać tego wzoru nie jest zbyt istotna – ważne jest, że taki wzór istnieje.

Nie jest to jedyne równanie diofantyczne, dla którego istnieje wzór dający wszystkie jego rozwiązania, ale przypadki takie są dość rzadkie. Należą do nich między innymi tak zwane równania Pella, takie jak $x^2 = 2y^2 + 1$. Równanie to ma nieskończenie wiele rozwiązań, na przykład: $3^2 = 2 \times 2^2 + 1$ czy $17^2 = 2 \times 12^2 + 1$, istnieje też odpowiedni ogólny wzór na rozwiązania. Trójki pitagorejskie mają jednak jeszcze głębszą strukturę, wywodzącą się również z geometrii. Załóżmy, że mamy dwie trójki pitagorejskie. W takim razie muszą istnieć dwa odpowiadające im rozwiązania równania Pitagorasa, które możemy przedstawić jako punkty wymierne na okręgu. Geometria pozwala nam w naturalny sposób „dodać” do siebie te punkty. Aby tego dokonać, znajdźmy punkt (1, 0), w którym okrąg przecina oś poziomą, i wyznaczmy dwa kąty między tym punktem i obydwoma rozwiązaniami. Dodajmy teraz te kąty do siebie (zob. ryc. 25) i sprawdźmy, jaki punkt w ten sposób otrzymaliśmy. Bez wątpienia punkt ten leży na okręgu, a po przeprowadzeniu krótkich obliczeń można się przekonać, że jest wymierny. Zatem w taki sposób możemy z dowolnych dwóch rozwiązań uzyskać trzecie. Matematycy wykryli już wiele takich związków. Większość z nich z miejsca wydaje się sensowna, gdy tylko pomyślimy o punktach wymiernych na okręgu.



Ryc. 25. Połączenie dwóch wymiernych rozwiązań A i B równania pitagorejskiego prowadzi do uzyskania trzeciego rozwiązania A + B

Przemycone przed chwilą „krótkie obliczenia” wymagają użycia trygonometrii. Klasyczne funkcje trygonometryczne, takie jak sinus i cosinus, są w sposób naturalny związane z geometrią okręgu. Przywołane tu obliczenia wykorzystują standardowe, dość eleganckie wzory wyrażające sinus i cosinus sumy dwóch kątów za pomocą sinusów i cosinusów kątów, które dodajemy. Istnieje wiele sposobów na obliczenie wartości funkcji sinus i cosinus, ale jeden z bardziej eleganckich polega na wykorzystaniu rachunku całkowego: wynik całkowania funkcji algebraicznej $1/\sqrt{1-x^2}$ można wyrazić za pomocą funkcji sinus. Dokładnie rzecz biorąc, chodzi tu o funkcję odwrotną do funkcji sinus – funkcję umożliwiającą wyznaczenie kąta na podstawie znanej wartości jego sinusa[51].

Całki pojawiają się wtedy, gdy próbujemy wyprowadzić wzór na długość łuku okręgu z wykorzystaniem rachunku różniczkowego i całkowego – okazuje się, geometria okręgu ma tu proste, ale bardzo ważne znaczenie. Obwód okręgu jednostkowego wynosi 2π,

a zatem jeśli przejdziemy po okręgu odległość 2π, to dojdziemy do punktu wyjścia. Tak samo będzie w przypadku całkowitych wielokrotności 2π – musimy jedynie pamiętać o przyjętej w matematyce konwencji, że liczby dodatnie odpowiadają ruchowi w kierunku zgodnym z ruchem wskazówek zegara, a ujemne – przemieszczaniu się w kierunku przeciwnym. Wynika z tego, że wartość funkcji sinus i cosinus danej liczby nie zmieni się, jeśli dodamy do niej całkowitą wielokrotność 2π. O takich funkcjach mówimy, że są okresowe; w tym przykładzie ich okres wynosi 2π.

W XVIII i XIX wieku specjaliści od analizy matematycznej odkryli uogólnienie tej całki, a wraz z nim wiele interesujących nowych funkcji, analogicznych do znanych wcześniej funkcji trygonometrycznych. Te nowe funkcje okazały się niezwykle ciekawe – są to funkcje okresowe, tak jak sinus i cosinus, ale w znacznie bardziej skomplikowany sposób. Nie mają jednego okresu – jak 2π (i całkowite wielokrotności tej wartości) w wypadku funkcji sinus i cosinus – ale dwa niezależne okresy. Gdybyśmy spróbowali znaleźć takie funkcje rzeczywiste, otrzymalibyśmy tylko stałe, ale dla funkcji zespolonych wachlarz możliwości jest znacznie bogatszy.

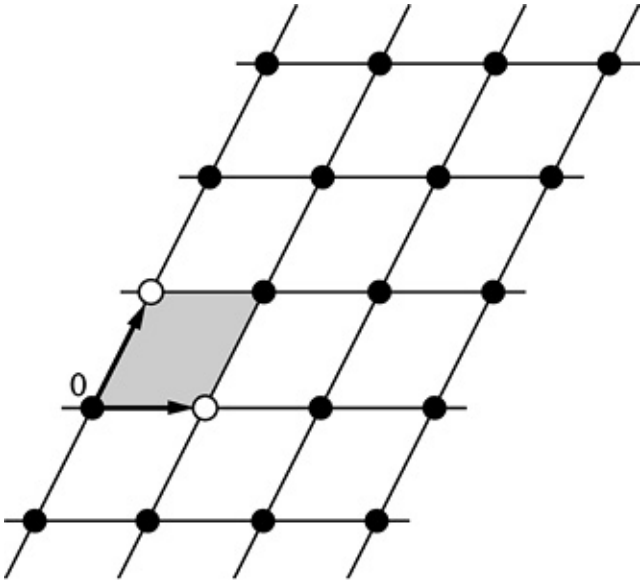
Badania w tym zakresie jako pierwsi rozpoczęli włoski matematyk Giulio di Fagnano i wielki Euler. Fagnano próbował wyznaczyć długość łuku elipsy za pomocą rachunku całkowego i różniczkowego, ale mimo usilnych starań nie udało mu się znaleźć odpowiedniego wzoru – obecnie nikogo to nie dziwi, ponieważ wiemy, że taki wzór nie istnieje. Zwrócił jednak przy okazji uwagę na związek między długościami różnych szczególnych łuków i swoje spostrzeżenia opisał w 1750 roku. Euler zauważył tę samą zależność w takim samym kontekście i przedstawił ją w postaci formalnego związku między całkami. Są to całki podobne do całki związanej z funkcją sinus, ale tu zamiast wyrażenia $1-x^2$ pod pierwiastkiem kwadratowym znajduje się wielomian trzeciego lub czwartego stopnia – może to być na przykład następujący wielomian czwartego stopnia: $\left(1-x^2\right)\left(1-4x^2\right)$.

W 1811 roku ukazał się pierwszy tom obszernego trzytomowego traktatu Adriena-Marie Legendre'a poświęconego tym całkom, które za sprawą związku z długością łuku elipsy nazywano całkami eliptycznymi. Legendre'owi umknęła jednak najważniejsza ich cecha: istnienie nowych funkcji analogicznych do sinus i cosinus, których funkcje *odwrotne* dają się w prosty sposób wyrazić za pomocą tych całek[52]. Gauss, Niels Henrik Abel i Carl Jacobi szybko to dostrzegli. Gauss, co dla niego typowe, zachował odkrycie dla siebie. Abel przesłał w 1826 roku do Francuskiej Akademii Nauk artykuł opisujący wnioski z jego badań, ale Cauchy, prezes Akademii, zapodział gdzieś rękopis i praca ukazała się dopiero w 1841 roku, dwanaście lat po tragicznej, przedwczesnej śmierci Abela spowodowanej chorobą płuc. Jednak w 1827 roku opublikowano inny artykuł Abela poświęcony tej samej tematyce. Jacobi poświęcił tym nowym „funkcjom eliptycznym" cały opasły tom, który ukazał się w 1829 roku i skierował analizę zespoloną na zupełnie nowe tory.

Powstała w ten sposób piękna struktura związanych ze sobą własności, analogicznych do własności funkcji trygonometrycznych. Związek, na który zwrócili wcześniej uwagę Fagnano i Euler, można było teraz przedstawić w postaci listy wzorów łączących funkcje eliptyczne sumy dwóch liczb z funkcjami eliptycznymi liczb będących składnikami tej sumy. Najwspanialsza cecha funkcji eliptycznych bije na głowę funkcje trygonometryczne: otóż funkcje eliptyczne nie dość, że są okresowe, to są okresowe podwójnie. Prosta jest jednowymiarowa, a więc wzdłuż prostej wzory mogą się powtarzać tylko w jednym kierunku. Płaszczyzna zespolona ma dwa wymiary, dzięki czemu wzory mogą się powtarzać tak samo jak na tapecie: w pionie, w dół ściany, ale także w bok, w kierunku sąsiednich pasów tapety. Z każdą funkcją eliptyczną związane są dwie niezależne liczby zespolone – jej okresy – i dodanie którejkolwiek z tych liczb do zmiennej nie zmienia wartości funkcji.

Powtarzając wcześniejsze rozważania, możemy dojść do wniosku, że wartość funkcji również się nie zmieni, jeśli dodamy do zmiennej dowolną kombinację całkowitą obydwu okresów. Kombinacje te mają interpretację geometryczną, a mianowicie wyznaczają one sieć na

płaszczyźnie zespolonej. Sieć taka dzieli płaszczyznę na równoległoboki i wszystko, co dzieje się na jednym z nich, jest powielane na pozostałych (zob. ryc. 26). Jeśli przyjrzymy się jednemu z tych równoległoboków i zwrócimy uwagę na to, jak łączy się z sąsiednimi polami, to zauważymy, że możemy utożsamić ze sobą jego przeciwległe krawędzie, podobnie jak w przypadku torusa utożsamialiśmy przeciwległe krawędzie kwadratu (zob. ryc. 12). Z topologicznego punktu widzenia równoległobok z utożsamionymi ze sobą przeciwległymi bokami również jest torusem. Zatem tak jak funkcje sinus i cosinus są związane z okręgiem, tak funkcje eliptyczne są związane z torusem.



Ryc. 26. Sieć na płaszczyźnie zespolonej. Strzałki wskazują na obie wartości okresu zaznaczone białymi kropkami. Wartości funkcji z obszaru równoległoboku zaznaczonego szarym kolorem powtarzają się na wszystkich pozostałych równoległobokach.

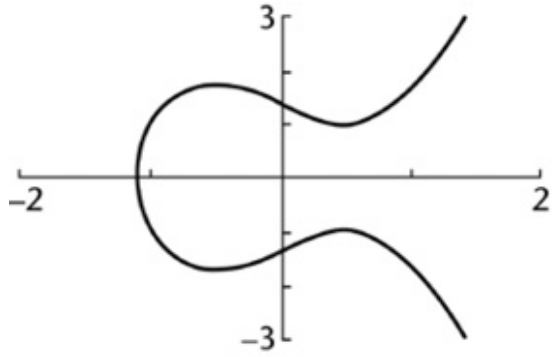
Istnieje również pewien związek z teorią liczb. Powiedzieliśmy wcześniej, że funkcję odwrotną do funkcji sinus można otrzymać, całkując wzór zawierający pierwiastek kwadratowy z wielomianu kwadratowego. Funkcje eliptyczne są podobne, ale wielomian kwadratowy należy tu zastąpić wielomianem trzeciego lub czwartego

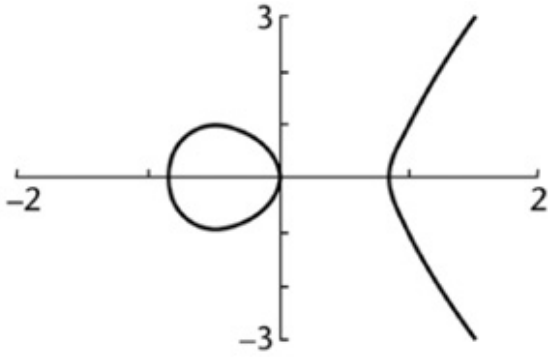
stopnia. Wcześniej podaliśmy przykład wielomianu czwartego stopnia, ponieważ chronologicznie tę zależność odkryto wcześniej, ale teraz skupmy uwagę na wielomianie sześciennym. Jeśli oznaczymy nasz pierwiastek kwadratowy symbolem *y*, a wielomian przedstawimy symbolicznie jako $ax^3 + bx^2 + cx + d$, gdzie *a*, *b*, *c* i *d* są współczynnikami liczbowymi, to *x* i *y* spełniają równanie:

$$y^2 = ax^3 + bx^2 + cx + d$$

Można je rozważać w kilku różnych kontekstach, w zależności od tego, jakie ograniczenia nałożymy na zmienne i współczynniki. Jeśli będą to wartości rzeczywiste, to równanie takie wyznacza krzywą na płaszczyźnie. Jeśli zaś użyjemy liczb zespolonych, to specjaliści od geometrii algebraicznej w dalszym ciągu będą nazywali zbiór rozwiązań tego równania krzywą, teraz jednak będzie to krzywa w przestrzeni par liczb zespolonych, która jest czterowymiarowa, jeśli wyrazi się ją za pomocą współrzędnych rzeczywistych. Ponadto z punktu widzenia liczb rzeczywistych krzywa ta jest w istocie powierzchnią.

Na rycinie 27 pokazano dwa typowe przykłady rzeczywistych krzywych eliptycznych: $y^2 = 4x^3 - 3x + 2$ oraz $y^2 = 4x^3 - 3x$. Ponieważ *y* występuje w równaniu podniesione do kwadratu, krzywe te są symetryczne względem osi poziomej. W zależności od współczynników wykres jest albo pojedynczą powyginaną krzywą, albo ma oddzielny element owalny. W dziedzinie liczb zespolonych krzywa eliptyczna jest zawsze w jednym, niepodzielonym kawałku.

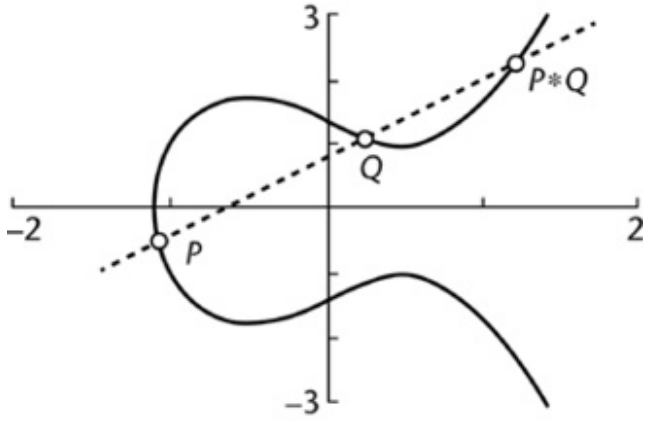




Ryc. 27. Typowe rzeczywiste funkcje eliptyczne. Po lewej: $y^2 = 4x^3 - 3x + 2$. *Po prawej:* $y^2 = 4x^3 - 3x$.

Teoria liczb dochodzi do głosu wtedy, gdy dopuszcza się, by zmienne i współczynniki były liczbami wymiernymi. W takim wypadku mamy przed sobą równanie diofantyczne. To dość mylące, że nazywa się je krzywymi eliptycznymi, mimo iż w niczym nie przypominają elipsy – przyczyną tego jest oczywiście związek z funkcjami eliptycznymi. To tak, jakbyśmy nazwali okrąg krzywą trójkątną, ponieważ ma związek z trygonometrią. Niestety taka nazwa już się przyjęła i musimy z tym jakoś żyć.

Ponieważ matematycy przeprowadzili szeroko zakrojone i bardzo owocne badania teoretyczne funkcji eliptycznych, specjalistom od teorii liczb udało się odkryć niezliczoną liczbę pięknych własności krzywych eliptycznych. Jedna z nich jest bliskim odpowiednikiem sposobu, w jaki możemy połączyć ze sobą dwa rozwiązania równania pitagorejskiego, dodając do siebie związane z nimi kąty. Jeśli chcemy powiązać ze sobą dwa punkty na krzywej eliptycznej, musimy narysować przebiegającą przez nie prostą i znaleźć punkt, w którym prosta ta przecina krzywą po raz trzeci (zob. ryc. 28). (Trzeci punkt przecięcia zawsze istnieje, ponieważ mamy do czynienia z równaniem trzeciego stopnia. Może się jednak zdarzyć, że znajduje się on „w nieskończoności” albo może być tożsamy z jednym z dwóch pierwszych punktów, jeśli prosta jest styczna do krzywej). Przyjmijmy, że dwa pierwsze punkty to *P* i *Q*, a trzeci oznaczmy jako $P * Q$.



Ryc. 28. Wyznaczanie punktu $P*Q$

Z obliczeń wynika, że jeśli *P* i *Q* są punktami wymiernymi, to również punkt $P*Q$ będzie wymierny. Działanie $*$ nadaje zbiorowi punktów wymiernych strukturę algebraiczną, ale okazuje się, że warto rozważyć inną związaną z nim operację. Wybierzmy dowolny punkt wymierny *O* leżący na krzywej i zdefiniujmy „dodawanie" punktów w następujący sposób:

$$P+Q=\left(P*Q\right)*O$$

To nowe działanie zachowuje podstawowe prawa tradycyjnej algebry (z tym że *O* odgrywa tu rolę zera) i zmienia zbiór wszystkich punktów wymiernych w strukturę, którą algebraicy nazywają grupą – więcej na ten temat powiemy w rozdziale 10. Najistotniejsze jest to, że podobnie jak w przypadku trójek pitagorejskich, możemy „dodać" do siebie dowolne dwa rozwiązania i uzyskać w ten sposób trzecie. Istnienie takiego „działania grupowego" w zbiorze punktów wymiernych jest uderzające i w szczególności oznacza, że jeśli tylko znajdziemy dwa rozwiązania równania diofantycznego, to automatycznie możemy ich uzyskać znacznie więcej.

Około roku 1908 Poincaré zastanawiał się, czy istnieje skończona liczba rozwiązań, która pozwala na uzyskanie wszystkich pozostałych przez wielokrotne zastosowanie działania grupowego. To ważne, ponieważ jeśli tak jest, to *wszystkie* rozwiązania wymierne można

opisać za pomocą skończonej ich liczby. We wspaniałym artykule z 1922 roku Mordell udowodnił, że odpowiedź na postawione przez Poincarégo pytanie brzmi: „tak". Tym samym krzywe eliptyczne zajęły kluczowe miejsce w teorii liczb, ponieważ posiadanie takiej kontroli nad równaniem diofantycznym jest czymś niezwykłym.

Zatem zarówno równanie pitagorejskie, jak i krzywe eliptyczne mają nieskończenie wiele rozwiązań wymiernych. W wypadku wielu równań diofantycznych sytuacja wygląda jednak inaczej – mają one tylko skończoną liczbę rozwiązań, a często nawet ani jednego. Pozwolę sobie teraz na krótką dygresję, żeby omówić całą rodzinę takich równań i niedawno przeprowadzony wspaniały dowód tego, że ich oczywiste rozwiązania są jedynymi, które istnieją.

Pitagorejczycy interesowali się równaniami, ponieważ wierzyli, że Wszechświat opiera się na liczbach. Na poparcie tego poglądu wskazywali, że harmonią w muzyce rządzą proste stosunki liczbowe. Przekonali się o tym, przeprowadziwszy doświadczenia z napiętą struną. Jeśli skrócimy strunę o połowę, nie zmieniając jej naprężenia, to uzyskamy dźwięk wyższy o oktawę. Jest to najbardziej harmonijne połączenie dwóch dźwięków – tak harmonijne, że brzmi nieco mdło. Kolejnymi co do znaczenia interwałami dźwięków w muzyce zachodniej są kwarta (jedna struna ma długość wynoszącą $\frac{3}{4}$ długości drugiej struny) oraz kwinta (w tym wypadku jedna struna ma $\frac{2}{3}$ długości drugiej)[53].

Jeśli wyjdziemy od liczby 1 i będziemy wielokrotnie mnożyć ją przez 2 lub 3, to otrzymamy liczby 2, 3, 4, 6, 8, 9, 12 i tak dalej – czyli liczby postaci 2^a3^b. Z uwagi na ich związek z muzyką nazywamy je liczbami harmonicznymi. W XIII wieku Gerszom ben Salomon Catalan, pisarz żydowski mieszkający we Francji, napisał książkę zatytułowaną *Sha'ar ha-Shamayim* (*Wrota niebios*), która była encyklopedią opartą na arabskich i greckich źródłach. Swoje dzieło podzielił na trzy części: fizyka, astronomia i metafizyka. W 1343 roku biskup Meaux namówił syna Gerszoma (a przynajmniej historycy uważają, że prawdopodobnie był to jego syn) Lewiego ben Gerszoma do napisania książki poświęconej matematyce,

zatytułowanej *Harmonia liczb*. Znalazło się w niej pytanie postawione przez kompozytora i teoretyka muzyki Philippe'a de Vitry'ego: ile jest par liczb harmonicznych różniących się o 1? Nietrudno znaleźć ich przykłady – de Vitry znał cztery takie pary: (1, 2), (2, 3), (3, 4) i (8, 9). Ben Gerszom dowiódł, że są to jedyne możliwe rozwiązania.

Spośród par de Vitry'ego najciekawsza jest (8, 9). Pierwsza liczba jest sześcianem: 2^3, druga zaś – kwadratem: 3^2. Matematycy zaczęli się zastanawiać, czy również inne kwadraty i sześciany mogą się różnić o 1, i Euler dowiódł, że to niemożliwe, z wyjątkiem trywialnego przypadku (0, 1), a jeśli dopuścimy całkowite liczby ujemne, to również (–1, 0). W 1844 roku drugi Catalan w naszej opowieści opublikował drukiem znacznie dalej idące twierdzenie, o którym zapewne rozmyślało wielu matematyków, ale nikomu nie przyszło do głowy, że warto je sformułować jawnie. Chodzi o belgijskiego matematyka Eugène'a Charles'a Catalana, który w 1844 roku przesłał do jednego z czołowych czasopism matematycznych, zatytułowanego „Journal für die Reine und Angewandte Mathematik" („Czasopismo poświęcone matematyce teoretycznej i stosowanej"), następujący list:

> Zwracam się z uprzejmą prośbą, żeby zechciał Pan opublikować w swoim czasopiśmie następujące twierdzenie, o którego prawdziwości jestem przekonany, choć nie udało mi się jej dotąd dowieść – być może inni odniosą większy sukces. Dwie kolejne liczby całkowite różne od 8 i 9 nie mogą być kolejnymi potęgami. Innymi słowy, równanie $x^m - y^n = 1$, w którym niewiadome są dodatnimi liczbami całkowitymi, ma tylko jedno rozwiązanie.

Stwierdzenie to stało się znane jako hipoteza Catalana. Wykładniki *m* i *n* są liczbami całkowitymi większymi od 1.

Mimo częściowych sukcesów hipoteza Catalana uparcie broniła się przed licznymi zakusami matematyków i zdołał ją udowodnić dopiero Preda Mihăilescu w 2002 roku. Mihăilescu urodził się w Rumunii w 1955 roku, a w 1973 przeprowadził się do Szwajcarii, gdzie

dopiero w 1997 roku uzyskał doktorat. Jego rozprawa doktorska, zatytułowana *Cyclotomy of rings and primality testing* (*Cyklotomia pierścieni a testy pierwszości*), była poświęcona zastosowaniom teorii liczb w testach pierwszości (mówiliśmy o nich w rozdziale 2). Problem ten nie ma szczególnego związku z hipotezą Catalana, ale Mihăilescu uświadomił sobie, że w przypadku metod, którymi się posługiwał, związek taki z pewnością istnieje. Bazują one na ideach, o których wspomnieliśmy w rozdziale 3, a mianowicie na gaussowskiej konstrukcji siedemnastokąta foremnego i związanych z nią równaniach algebraicznych. Rozwiązania tych równań nazywane są liczbami cyklotomicznymi. Dowód Mihăilescu jest bardzo skomplikowany i był prawdziwym szokiem dla całej społeczności matematyków. Wynika z niego, że bez względu na to, jakie wartości wybierzemy dla obydwu potęg, liczba rozwiązań jest skończona i poza oczywistymi rozwiązaniami wykorzystującymi liczby 0 i ±1 jedynym interesującym przypadkiem jest $3^2 - 2^3 = 1$.

Powyższe przykłady pokazują, że niektóre równania diofantyczne mają nieskończenie wiele rozwiązań, a inne – ani jednego. No i co z tego? – można by zapytać. Przecież te przypadki obejmują wszystkie możliwości. Jeśli jednak zaczniemy się zastanawiać, które równania są jakiego typu, to sytuacja robi się ciekawsza. W czasach gdy Mordell, ekspert od równań diofantycznych, pisał swój jakże ważny podręcznik, cała ta dziedzina badań przypominała biologię w jej początkowym okresie – mnóstwo zebranych motyli i niewiele postępów na drodze do ich systematyzacji. Oto kolorowy paź pitagorejski, tu widać wielką rusałkę eliptyczną, a tam w krzakach znajdziemy nakrapiane gąsienice bielinka Pella. Cała dziedzina była ciągle w takim stanie, w jakim zostawił ją Diofantos, zwiększyła się tylko liczba okazów – uczeni zgromadzili nieuporządkowaną listę niezależnych metod, po jednej dla każdego rodzaju równania. Z takiego materiału nie da się stworzyć dobrego podręcznika. Trzeba to wszystko najpierw uporządkować – Mordell zabrał się więc do roboty.

W pewnym momencie musiał zauważyć, że wszystkie równania, o których wiadomo, że mają nieskończenie wiele rozwiązań wymiernych – takie jak równanie pitagorejskie i krzywe eliptyczne – charakteryzuje jedna wspólna cecha. Mordell skupił się na jednej klasie równań: na równaniach, w których (po przekształceniu ich do postaci z liczbami wymiernymi, tak jak to zrobiliśmy z równaniem Pitagorasa) występują tylko dwie zmienne. Istnieją dwa typy równań, dla których wiemy, jak znaleźć nieskończenie wiele rozwiązań. Przykładem pierwszego z nich jest równanie pitagorejskie zapisane w równoważnej postaci: $x^2 + y^2 = 1$. W tym wypadku istnieje wzór na uzyskanie wszystkich rozwiązań. Wystarczy do niego wstawić dowolną liczbę wymierną, by uzyskać jedno wymierne rozwiązanie, a z tego już wynikają wszystkie rozwiązania. Przykładem drugiego typu są krzywe eliptyczne: tu mamy *proces* pozwalający uzyskać nowe rozwiązania z rozwiązań już poznanych i gwarancję, że jeśli znajdziemy na początku odpowiedni skończony ich zbiór, to proces ten pozwoli uzyskać wszystkie możliwe rozwiązania.

Hipoteza Mordella głosi, że w każdym przypadku, w którym mamy nieskończenie wiele rozwiązań, musimy mieć do czynienia z jedną z tych dwóch opcji. Albo musi istnieć ogólny wzór, albo proces pozwalający uzyskać wszystkie rozwiązania z odpowiedniego, skończonego ich zbioru. We wszystkich pozostałych sytuacjach liczba rozwiązań wymiernych jest skończona – przykładem tego mogą być równania $x^m - y^n = 1$, występujące w hipotezie Catalana. W pewnym sensie rozwiązania są wtedy jedynie dziełem przypadku i są pozbawione jakiejkolwiek struktury podstawowej.

Mordell doszedł do tego wniosku nieco inną drogą. Zauważył, że każde równanie o nieskończenie wielu rozwiązaniach wymiernych ma pewną uderzającą cechę topologiczną, a mianowicie ma genus równy 0 lub 1. Jak pamiętamy z rozdziału 4, genus jest pojęciem z dziedziny topologii i mówi o tym, ile w danej powierzchni jest otworów. Sfera ma genus równy zero, genus torusa wynosi 1, dla torusa z dwoma otworami genus jest równy 2 i tak dalej. Skąd się wzięły powierzchnie w rozważaniach na temat problemu z zakresu

teorii liczb? Z geometrii analitycznej. Mówiliśmy o tym przy okazji równania pitagorejskiego rozważanego w dziedzinie liczb wymiernych, które po rozszerzeniu na liczby rzeczywiste opisuje okrąg. Mordell zrobił kolejny krok i dopuścił, żeby rozwiązania były również liczbami zespolonymi. Każde równanie z dwiema zmiennymi zespolonymi opisuje tak zwaną krzywą zespoloną. Jednak z punktu widzenia liczb rzeczywistych i naszego sposobu ich obrazowania każda liczba zespolona jest dwuwymiarowa: ma dwie składowe – część rzeczywistą i urojoną – wyrażane za pomocą liczb rzeczywistych. Zatem „krzywa" na powierzchni zespolonej jest dla nas, zwykłych śmiertelników, powierzchnią. A skoro tak, to ma jakąś wartość genusu – no i wszystko jasne.

Okazało się, że każde równanie, o którym było wiadomo, że ma tylko skończoną liczbę rozwiązań, ma genus o wartości przynajmniej 2. Ważne równania, których charakter pozostawał niewyjaśniony, również miały genus równy 2 lub więcej. Zdobywając się na szalony i odważny krok, poparty dość wątłymi, jak się wówczas wydawało, dowodami, Mordell wysunął hipotezę, że każde równanie diofantyczne o genusie równym przynajmniej 2 ma skończoną liczbę rozwiązań. Tym jednym posunięciem udało mu się podzielić diofantyczne motyle na spokrewnione ze sobą rodziny.

Hipoteza Mordella miała tylko jeden niewielki szkopuł. Łączyła ze sobą dwie zupełnie różne rzeczy: wymierne rozwiązania równań i topologię. W czasach Mordella taki związek wydawał się niezwykle mało prawdopodobny. Nawet jeśli w ogóle istniał, to i tak nikt nie miał pojęcia, jak go odkryć. Zatem hipoteza Mordella wyglądała na szalone, nieuzasadnione domniemanie, ale potencjalne korzyści, jakie mogła przynieść, były ogromne.

W 1983 roku Gerd Faltings opublikował wspaniały dowód pokazujący, że śmiała hipoteza Mordella jest w istocie prawdziwa. W dowodzie wykorzystał zaawansowane metody geometrii algebraicznej. Niedługo potem Paul Vojta przedstawił zupełnie inny dowód, wykorzystujący przybliżenia liczb rzeczywistych za pomocą liczb wymiernych, a w 1990 roku Enrico Bombieri opublikował uproszczony dowód oparty na tej samej metodzie. Twierdzenie Faltingsa znalazło zastosowanie w dowodzie wielkiego twierdzenia

Fermata, o którym szerzej powiemy w rozdziale 7. Chodzi o to, że dla każdej liczby całkowitej *n* większej lub równej 3 równanie $x^n + y^n = 1$ ma najwyżej skończoną liczbę rozwiązań. Genus krzywej związanej z tym równaniem ma wartość $(n-1)(n-2)/2$, która wynosi przynajmniej 3 dla *n* równego lub większego od 4. Z twierdzenia Faltingsa od razu wynika, że dla każdego $n \geq 4$ równanie Fermata ma najwyżej skończoną liczbę rozwiązań wymiernych. Fermat twierdził, że w ogóle nie ma rozwiązań z wyjątkiem przypadku, gdy *x* lub *y* jest równe zero, był to więc duży krok naprzód. W następnym rozdziale powrócimy do opowieści o wielkim twierdzeniu Fermata i pokażemy, jak udało się je w pełni udowodnić.

50 Andrew Granville, Thomas Tucker, *It's as easy as* abc (*To proste jak abecadło*), „Notices of the American Mathematical Society" 2002, tom 49, s. 1224–1231.

51 Mówiąc dokładnie, chodzi o wzór:

$$\int \frac{dx}{\sqrt{1-x^2}} = \arcsin x$$

gdzie arcsin (często zapisywany jako $\sin^{-1}$) jest funkcją odwrotną do funkcji sinus. Innymi słowy, jeśli $y = \sin x$, to $x = \arcsin y$.

52 Załóżmy na przykład, że *k* jest dowolną liczbą zespoloną, i rozważmy następującą całkę:

$$\int \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}}$$

Jest to funkcja odwrotna do funkcji eliptycznej oznaczanej symbolicznie jako sn. Dla każdej wartości *k* istnieje jedna taka funkcja. Jest to relacja podobna do omawianej w poprzedniej uwadze (nr 51), ale bardziej złożona.

53 Zob.: Ian Stewart, *17 równań, które zmieniły świat*, przeł. Julia Szajkowska, Prószyński i S-ka, Warszawa 2013, rozdział 8.

7. Zbyt wąskie marginesy

Wielkie twierdzenie Fermata

Z Fermatem spotkaliśmy się już w rozdziale 2, w którym jego eleganckie twierdzenie o potęgach liczb pozwoliło nam znaleźć metodę sprawdzania, czy dana liczba jest liczbą pierwszą. Ten rozdział poświęcimy na omówienie dużo trudniejszego, wielkiego twierdzenia Fermata, nazywanego również jego ostatnim twierdzeniem. Brzmi to bardzo tajemniczo. Słowo „twierdzenie" jest jasne, ale kim był Fermat i dlaczego było to jego *ostatnie* twierdzenie? Czy ta nazwa jest tylko jakimś sprytnym chwytem marketingowym? Okazuje się, że nie: taką nazwę nadano temu problemowi w XVIII wieku, gdy tylko kilku najwybitniejszych matematyków wiedziało o jego istnieniu i się nim interesowało. Nie ulega jednak wątpliwości, że wielkie twierdzenie Fermata jest naprawdę tajemnicze.

Pierre Fermat urodził się we Francji, według jednych źródeł w 1601, a według innych w 1607 lub 1608 roku. Ta niepewność być może wynika z tego, że miał brata o takim samym imieniu. Jego ojciec był bogatym kupcem handlującym wyrobami skórzanymi i zajmował wysokie stanowisko w lokalnych władzach, natomiast matka wywodziła się z rodziny prawniczej. Uczył się na uniwersytecie w Tuluzie, a pod koniec lat dwudziestych XVII wieku przeniósł się do Bordeaux, gdzie ujawnił się jego talent matematyczny. Władał biegle kilkoma językami i zajął się odtworzeniem zaginionych prac greckiego matematyka Apoloniusza. Swoimi licznymi odkryciami dzielił się z czołowymi matematykami owych czasów.

W 1631 roku, po uzyskaniu stopnia naukowego w dziedzinie prawa na uniwersytecie w Orleanie, Fermat został mianowany radcą Izby Karnej Sądu Najwyższego w Tuluzie. Uprawniało go to do dodania do nazwiska arystokratycznego przyimka „de". Pozostał już na tym stanowisku do końca życia, ale jego prawdziwą pasją była matematyka. Publikował niewiele – wolał opisywać swoje odkrycia w listach, zwykle nie podając dowodu. Jego prace cieszyły się dużym uznaniem wśród wybitnych matematyków i z wieloma z nich był po imieniu, choć wciąż pozostawał tylko amatorem. Fermat okazał się jednak tak utalentowany, że w istocie był prawdziwym matematykiem – nie miał tylko oficjalnego tytułu naukowego.

Niektóre z jego dowodów przetrwały w listach i artykułach. Bardzo wyraźnie na ich przykładzie widać, że Fermat dobrze wiedział, jak wygląda prawdziwy dowód. Po jego śmierci wiele z jego najgłębszych twierdzeń pozostało bez dowodu i matematycy zabrali się do pracy. Po kilku dziesięcioleciach już tylko jedno twierdzenie Fermata nie miało dowodu, w naturalny sposób zaczęto je więc nazywać ostatnim twierdzeniem Fermata. W przeciwieństwie do pozostałych, nikt nie mógł się z nim uporać i jaskrawy kontrast między prostotą jego sformułowania a olbrzymimi trudnościami ze znalezieniem dowodu sprawił, że szybko zrobiło się o nim głośno.

Wydaje się, że Fermat sformułował swoje wielkie twierdzenie około roku 1630. Dokładna data nie jest znana, ale było to w okresie, gdy zaczął czytać wydaną niedawno *Arytmetykę* Diofantosa. Właśnie podczas tej lektury wpadł na pomysł swojego twierdzenia. Wielkie twierdzenie ukazało się po raz pierwszy drukiem w 1670 roku, pięć lat po śmierci Fermata, gdy jego syn Samuel opublikował nowe wydanie *Arytmetyki*. Wydanie to wyróżniało się tym, że znalazły się w nim notatki Pierre'a de Fermata, które zapisał na marginesach własnego egzemplarza tego dzieła – był to łaciński przekład z 1621 roku autorstwa Claude'a Gasparda Bacheta de Méziriaca. Fermat zapisał wielkie twierdzenie w postaci uwagi do zadania VIII z księgi II (zob. ryc. 29).

QVÆSTIO VIII.

PROPOSITVM quadratum diuidere in duos quadratos. Imperatum sit vt 16. diuidatur in duos quadratos. Ponatur primus 1 Q. Oportet igitur 16 − 1 Q. æquales esse quadrato. Fingo quadratum à numeris quotquot libuerit, cum defectu tot vnitatum quod continet latus ipsius 16. esto à 2 N. − 4. ipse igitur quadratus erit 4 Q. + 16. − 16 N. hæc æquabuntur vnitatibus 16 − 1 Q. Communis adiiciatur vtrimque defectus, & à similibus auferantur similia, fient 5 Q. æquales 16 N. & fit 1 N. $\frac{16}{5}$ Erit igitur alter quadratorum $\frac{256}{25}$. alter verò $\frac{144}{25}$ & vtriusque summa est $\frac{400}{25}$ seu 16. & vterque quadratus est.

Τὸν δοθέντα τετράγωνον διελεῖν εἰς δύο τετραγώνους. ἐπιτετάχθω δὴ τὸ ις̄ διελεῖν εἰς δύο τετραγώνους. καὶ τετάχθω ὁ πρῶτος δυνάμεως μιᾶς. δεήσει ἄρα μονάδας ις̄ λείψει δυνάμεως μιᾶς ἴσας ϛ̄ τετραγώνῳ. πλάσω τὸ τετράγωνον ἀπὸ ς̄ς̄. ὅσων δή ποτε λείψει τοσούτων μ̄ ὅσων ἔχει ἡ πλευρὰ ἴσον ἔσται ς̄ς̄ β̄ λείψει μ̄ δ̄. αὐτὸς ἄρα ὁ τετράγωνος ἔσται δυνάμεων δ̄ μ̄ ις̄ λείψει ς̄ς̄ ις̄. ταῦτα ἴσα μονάσι ις̄ λείψει δυνάμεως μιᾶς. κοινὴ προσκείσθω ἡ λεῖψις, καὶ ἀπὸ ὁμοίων ὁμοῖα. δυνάμεις ἄρα ε̄ ἴσαι ἀριθμοῖς ις̄. καὶ γίνεται ὁ ἀριθμὸς ις̄. πέμπτων. ἔσται ὁ μβ̄ σνς̄ εἰκοσοπέμπτων. ὁ δὲ ρμδ̄ εἰκοσοπέμπτων, ἐστὶ δὲ καὶ οἱ δύο συντιθέντες ποιοῦσι τὸν εἰκοσόπεμπτα, ἕνι μονάδας ις̄. καὶ ἔστιν ἑκάτερος τετράγωνος.

OBSERVATIO DOMINI PETRI DE FERMAT.

CVbum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos & generaliter nullam in infinitum vltra quadratum potestatem in duos eiusdem nominis fas est diuidere cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

Ryc. 29. Notatka zapisana przez Fermata na marginesie, która ukazała się w wydaniu Arytmetyki *Diofantosa przygotowanym przez jego syna, Samuela*

Diofantos rozwiązuje w tym miejscu zadanie, w którym należy zapisać kwadrat liczby jako sumę kwadratów dwóch innych liczb. W rozdziale 6 przekonaliśmy się, że istnieje nieskończenie wiele takich trójek pitagorejskich. Diofantos stawia podobne, ale trudniejsze pytanie: w jaki sposób można znaleźć dwa krótsze boki trójkąta prostokątnego, jeśli znamy długość jego najdłuższego boku? Podany kwadrat trzeba „rozdzielić” na dwa kwadraty, czyli wyrazić go za pomocą ich sumy. Diofantos pokazuje, jak należy rozwiązać takie zadanie, gdy przeciwprostokątna ma długość 4, i uzyskuje odpowiedź w zbiorze liczb wymiernych

$$4^2 = (16/5)^2 + (12/5)^2$$

Po pomnożeniu obu stron równania przez 25 otrzymujemy: $20^2 = 16^2 + 12^2$, a po podzieleniu przez 16 dostajemy znaną równość $3^2 + 4^2 = 5^2$. Postępując zgodnie z tradycją sięgającą starożytnego Babilonu, Diofantos ilustruje zwykle ogólne metody konkretnymi przykładami i nie przedstawia dowodów.

Egzemplarz *Arytmetyki* należący do Fermata nie dotrwał do naszych czasów, ale Fermat musiał w nim zapisać swoją uwagę na marginesie, bo tak twierdzi Samuel. Zapewne często zaglądał do owej cennej książki i jego hipoteza jest tak naturalna, że prawdopodobnie przyszła mu na myśl od razu, gdy tylko przeczytał zadanie VIII z księgi II. Najwyraźniej zastanawiał się, czy w podobny sposób można by postąpić w przypadku sześcianów – to zupełnie naturalne pytanie dla każdego matematyka. Nie znalazł żadnych takich przykładów – możemy być tego pewni, bo to niemożliwe – i tak samo bezowocnie zakończyły się jego poszukiwania tego typu równości dla wyższych potęg, na przykład dla potęgi czwartej. Doszedł do wniosku, że takie zadania nie mają rozwiązania. Tego właśnie można się dowiedzieć z zapisanej na marginesie notatki. W przekładzie na język polski brzmi ona następująco:

> Nie można podzielić sześcianu na dwa sześciany, ani czwartej potęgi na dwie czwarte potęgi, ani w ogóle żadnej potęgi wyższej od drugiej na dwie takie same potęgi. Odkryłem naprawdę wspaniały dowód tego stwierdzenia, ale margines jest zbyt wąski, by się tu zmieścił.

Mówiąc językiem algebry, Fermat twierdzi, że udało mu się udowodnić, iż równanie diofantyczne

$$x^n + y^n = z^n$$

nie ma rozwiązań w zbiorze liczb naturalnych, gdy *n* jest liczbą całkowitą większą lub równą 3. Oczywiście pomija tu trywialne rozwiązania, w których *x* lub *y* jest równe zero. Aby uniknąć konieczności częstego powtarzania tego wzoru, nazwijmy go równaniem Fermata.

Jeśli Fermat faktycznie miał taki dowód, to nikomu nie udało się go odkryć. Twierdzenie udowodniono ostatecznie w 1995 roku, ponad 350 lat po jego sformułowaniu, ale metody wykorzystane w tym dowodzie wykraczają daleko poza techniki dostępne w czasach Fermata i wszystko, co mógł wówczas sam wymyślić. Poszukiwania tego dowodu wywarły olbrzymi wpływ na rozwój matematyki. W zasadzie były przyczyną powstania algebraicznej teorii liczb. Dziedzina ta rozwinęła się bujnie w XIX stuleciu za sprawą nieudanej próby udowodnienia wielkiego twierdzenia Fermata i genialnej idei, która pozwoliła częściowo uratować ten nieudany dowód. Natomiast pod koniec XX i na początku XXI wieku prace nad dowodem twierdzenia Fermata wywołały prawdziwą rewolucję.

Pierwsi matematycy, którzy zajęli się twierdzeniem Fermata, próbowali rozprawić się kolejno z poszczególnymi potęgami. Nie wiemy, czy Fermat rzeczywiście znalazł ogólny dowód, o którym wspomina w swojej notatce, ale wiemy, jak udowodnił to twierdzenie dla czwartej potęgi. Głównym narzędziem jest w tym wypadku przepis Euklidesa na trójki pitagorejskie. Czwarta potęga dowolnej liczby jest kwadratem jej kwadratu. Zatem każde rozwiązanie równania Fermata dla czwartej potęgi musi być trójką pitagorejską, w której wszystkie trzy liczby same są kwadratami. Ten dodatkowy warunek można wstawić do przepisu Euklidesa i po kilku sprytnych przekształceniach uzyskać *inne* rozwiązanie równania Fermata dla potęgi czwartej[54]. Mogłoby się wydawać, że to żaden postęp – po dokonaniu przekształceń algebraicznych zajmujących całą stronę udaje nam się jedynie zamienić jeden problem na drugi, praktycznie taki sam. Te przekształcenia doprowadziły jednak do zredukowania problemu: w drugim rozwiązaniu liczby są mniejsze od tych w pierwszym, hipotetycznym. Co istotne, jeśli pierwsze rozwiązanie

nie jest trywialne – to znaczy, jeśli *x* i *y* są różne od zera – to tak samo będzie w drugim. Fermat zauważył, że wielokrotne powtórzenie tej procedury pozwala na uzyskanie ciągu rozwiązań, w których liczby będą coraz mniejsze. Jednak każdy malejący ciąg liczb naturalnych musi się kiedyś skończyć. Doszliśmy więc do logicznej sprzeczności, a to oznacza, że hipotetyczne rozwiązanie nie istnieje. Fermat nazwał takie podejście „metodą nieskończonego schodzenia". Obecnie uznajemy ją za jeden z rodzajów dowodzenia indukcyjnego, o którym wspomnieliśmy w rozdziale 4. Metodę Fermata można wyrazić za pomocą pojęcia minimalnego kontrprzykładu, a raczej, w tym wypadku, minimalnego rozwiązania. Przypuśćmy, że istnieje jakieś nietrywialne rozwiązanie naszego równania. Jeśli tak, to musi również istnieć rozwiązanie minimalne. Jednak ze spostrzeżenia Fermata wynika, że musi w takim razie także istnieć jeszcze mniejsze rozwiązanie – i tak dochodzimy do sprzeczności. Oznacza to więc, że nie może istnieć żadne rozwiązanie. Później pojawiło się jeszcze wiele innych dowodów dla potęgi czwartej – obecnie znamy ich około 30.

Fermat wykorzystał prosty fakt, że czwarta potęga jest szczególnym przypadkiem kwadratu. Dzięki temu można zauważyć, że dowód wielkiego twierdzenia Fermata da się przeprowadzić dla dwóch przypadków, gdy potęga *n* jest albo 4, albo nieparzystą liczbą pierwszą. Każda liczba *n* większa od 2 dzieli się przez 4 lub przez nieparzystą liczbę pierwszą *p*, a zatem każda *n*-ta potęga jest potęgą albo czwartą, albo *p*-tą. W ciągu następnych dwóch stuleci udowodniono wielkie twierdzenie Fermata dla trzech nieparzystych liczb pierwszych: 3, 5, i 7. Euler zajął się sześcianami w 1770 roku i choć w opublikowanym dowodzie jest pewna luka, można ją zapełnić, wykorzystując wynik z innego artykułu Eulera. Legendre i Peter Dirichlet zajęli się piątą potęgą około roku 1825. Gabriel Lamé udowodnił wielkie twierdzenie Fermata dla siódmej potęgi w 1839 roku. Później matematycy znaleźli jeszcze wiele innych dowodów dla tych przypadków. Tymczasem kilku uczonych przedstawiło dowody dla potęg 6, 10 i 14, ale ostatecznie okazały się one zbędne, zastąpiły je bowiem dowody dla potęg 3, 5 i 7.

Każdy z tych dowodów wykorzystuje konkretne własności algebraiczne charakterystyczne dla rozważanych potęg. Nie ma w nich śladu żadnej ogólnej struktury, którą można by zastosować do przeprowadzenia dowodu dla wszystkich potęg, czy choćby dla większej ich liczby. Dowody dla większych potęg okazały się coraz bardziej skomplikowane. Potrzebna była jakaś nowa koncepcja, coś, co pozwoli spojrzeć na ten problem z nowej perspektywy. Sophie Germain, wybitna matematyczka, podzieliła wielkie twierdzenie Fermata dla potęgi *p*-tej będącej liczbą pierwszą na dwie kategorie. Do pierwszej zaliczyła te równości, w których żadna z liczb *x*, *y*, *z* nie jest podzielna przez *p*. Do drugiej te, w których jedna z tych liczb dzieli się przez *p*. Rozważając specjalne „pomocnicze" liczby pierwsze związane z *p,* Germain udowodniła, że w pierwszym wypadku twierdzenie Fermata nie ma rozwiązań dla potęg będących liczbami pierwszymi mniejszymi od 100. Okazało się jednak, że udowodnienie ogólnego twierdzenia z wykorzystaniem pomocniczych liczb pierwszych jest bardzo trudne.

Germain prowadziła korespondencję z Gaussem, używając początkowo męskiego pseudonimu, i wielki uczony był pod dużym wrażeniem jej pomysłowości. Gdy w końcu wyjawiła, że jest kobietą, jego podziw wzrósł jeszcze bardziej i otwarcie jej o tym powiedział. W przeciwieństwie do wielu jemu współczesnych Gauss nie zakładał, że kobiety nie mogą osiągać sukcesów w naukach teoretycznych, takich jak badania matematyczne. Później Germain przeprowadziła nieudaną próbę udowodnienia wielkiego twierdzenia Fermata pierwszej kategorii dla wszystkich potęg parzystych, a więc dla przypadku, gdy można wykorzystać Euklidesowy opis trójek pitagorejskich. Ostatecznie z potęgami parzystymi rozprawił się Guy Terjanian w 1977 roku. Druga kategoria okazała się dużo twardszym orzechem do zgryzienia i nikomu nie udało się zbyt wiele osiągnąć na tym polu.

W 1847 roku, po przeprowadzeniu dowodu dla siódmej potęgi, Lamé wpadł na wspaniały pomysł. Wymagał on wprowadzenia liczb zespolonych, ale w tamtych czasach nikomu to już nie

przeszkadzało. Sedno tego podejścia zasadzało się na tym samym pomyśle, który Gauss zastosował do udowodnienia możliwości przeprowadzenia konstrukcji siedemnastokąta foremnego (mówiliśmy o tym w rozdziale 3). Każdy specjalista od teorii liczb wiedział o jego istnieniu, ale dopiero Lamé zaczął się poważnie zastanawiać, czy mógłby on pomóc w udowodnieniu wielkiego twierdzenia Fermata.

W zbiorze liczb rzeczywistych jedynka ma dokładnie jeden pierwiastek *p*-tego stopnia (gdy *p* jest liczbą nieparzystą), a mianowicie jest nim sama liczba 1. Jednak w zbiorze liczb zespolonych 1 ma wiele pierwiastków *p*-tego stopnia, a mówiąc ściśle, jest ich dokładnie *p*. Wynika to z podstawowego twierdzenia algebry, ponieważ pierwiastki te są rozwiązaniem równania $x^p - 1 = 0$, które ma stopień *p*-ty. Istnieje elegancki wzór na zespolone pierwiastki *p*-tego stopnia z jedynki[55], który pokazuje, że są one potęgami 1, ζ, ζ^2, ζ^3, …, ζ^{p-1} konkretnej liczby zespolonej ζ. Z podstawowej własności tych liczb wynika, że równanie $x^p + y^p$ (dla *p* będącego liczbą nieparzystą) rozkłada się na *p* czynników:

$$x^p + y^p = (x+y)(x+\zeta y)(x+\zeta^2 y)\cdots(x+\zeta^{p-1} y)$$

W równaniu Fermata wyrażenie to jest równe z^p, czyli jest *p*-tą potęgą pewnej liczby całkowitej. Nietrudno jednak zauważyć, że jeśli iloczyn liczb niemających żadnego wspólnego czynnika jest *p*-tą potęgą, to każda z tych liczb sama musi być *p*-tą potęgą. Zatem po przeprowadzeniu pewnych przekształceń Lamé mógł zapisać każdy czynnik w postaci *p*-tej potęgi. Z tego zaś doszedł do sprzeczności.

Lamé ogłosił swój dowód wielkiego twierdzenia Fermata na forum Francuskiej Akademii Nauk w Paryżu w marcu 1847 roku, nadmieniając, że opiera się on na pomyśle Josepha Liouville'a. Liouville podziękował Lamému, ale wskazał na potencjalny problem. Otóż prawdziwość kluczowego stwierdzenia, z którego wynika, że każdy czynnik jest *p*-tą potęgą, nie jest wcale przesądzona. Zależy bowiem od tego, czy rozkład na czynniki pierwsze jest jednoznaczny – nie

tylko dla zwyczajnych liczb całkowitych, dla których jest to akurat prawdą, ale dla nowego rodzaju liczb, których użył tu Lamé. Chodzi o kombinacje potęg ζ, nazywane liczbami cyklotomicznymi całkowitymi. Słowo „cyklotomiczny" znaczy „dzielący okrąg" i odwołuje się do związku wykorzystanego przez Gaussa. Liouville stwierdził, że problem nie polega jedynie na braku dowodu jednoznaczności rozkładu na czynniki pierwsze liczb cyklotomicznych całkowitych – wcale nie jest wykluczone, że taki rozkład nie jest jednoznaczny.

Inni matematycy już wcześniej zgłaszali podobne wątpliwości. Trzy lata wcześniej Gotthold Eisenstein w jednym z listów napisał:

> Gdybyśmy mieli twierdzenie głoszące, że iloczyn dwóch liczb zespolonych jest podzielny przez liczbę pierwszą tylko wtedy, gdy dzieli się przez nią jeden z czynników – co wydaje się zupełnie oczywiste – to za jednym zamachem mielibyśmy całą teorię [liczb algebraicznych]; jednak takie twierdzenie jest całkowicie fałszywe.

Twierdzenie wspomniane w tym liście jest głównym krokiem w dowodzie jednoznaczności rozkładu na czynniki pierwsze. Eisenstein rozważał nie tylko liczby użyte przez Lamégo, ale także podobne liczby wynikające z innych równań. Są to tak zwane liczby algebraiczne. Liczbą algebraiczną nazywamy liczbę zespoloną będącą pierwiastkiem wielomianu o wymiernych współczynnikach. Natomiast liczbą algebraiczną całkowitą nazywamy liczbę zespoloną będącą pierwiastkiem wielomianu o współczynnikach całkowitych, przy założeniu, że współczynnik przy najwyższej potędze x jest równy 1. Dla każdego takiego wielomianu otrzymujemy związane z nim ciało liczb algebraicznych (co znaczy, że w wyniku dodawania, odejmowania, mnożenia oraz dzielenia takich liczb uzyskamy liczby tego samego rodzaju) i jego pierścień liczb algebraicznych całkowitych (tak samo jak poprzednio, tylko bez dzielenia). Są to podstawowe obiekty, które bada się w algebraicznej teorii liczb.

Jeśli przyjmiemy na przykład, że nasz wielomian ma postać x^2-2, to jego pierwiastkiem jest $\sqrt{2}$. Związane z nim ciało składa się ze wszystkich liczb postaci $a+b\sqrt{2}$, gdzie *a* i *b* są liczbami wymiernymi, pierścień liczb całkowitych zaś tworzą wszystkie liczby takiej samej postaci, ale dla *a* i *b* będących liczbami całkowitymi. Możemy zatem zdefiniować czynniki pierwsze i rozkład na nie jest jednoznaczny. Zdarzają się jednak pewne niespodzianki: wielomian x^2+x-1 ma pierwiastek $\left(\sqrt{5}-1\right)\Big/2$, a więc mimo ułamka jest to liczba algebraiczna *całkowita*.

W algebraicznej teorii liczb trudność nie polega na zdefiniowaniu czynników. Na przykład liczba cyklotomiczna całkowita jest czynnikiem innej liczby (czyli innymi słowy, jest jej dzielnikiem), jeśli ta druga jest równa pierwszej pomnożonej przez jakąś liczbę cyklotomiczną całkowitą. Trudność nie polega tu też na zdefiniowaniu liczb pierwszych: liczba cyklotomiczna całkowita jest liczbą pierwszą, jeśli nie ma czynników innych niż trywialne „wartości jednostkowe", czyli liczby cyklotomiczne całkowite będące dzielnikami 1. Nie mamy też kłopotów z rozkładem cyklotomicznej liczby całkowitej ani żadnej innej liczby algebraicznej na czynniki pierwsze. Wystarczy dzielić ją tak długo, jak się da, przez kolejne czynniki. Można prosto udowodnić, że taka procedura musi kiedyś dobiec końca, a gdy tak się stanie, każdy z czynników powinien być czynnikiem pierwszym. W czym zatem problem? W jednoznaczności. Jeśli przeprowadzimy taki rozkład ponownie, dokonując jednak po drodze innych wyborów, to możemy otrzymać na końcu inną listę czynników pierwszych.

Gdy rozważamy to po raz pierwszy, trudno nam zrozumieć, jak mogłoby do tego dojść. Czynniki pierwsze są najmniejszymi możliwymi elementami, na które można podzielić daną liczbę. Przypomina to rozkładanie konstrukcji z klocków Lego na pojedyncze elementy. Gdyby można było to zrobić na różne sposoby, oznaczałoby to, że jeden z klocków daje się rozłożyć na dwie lub więcej części, jednak świadczyłoby, że nie jest to pojedynczy klocek. Niestety, porównanie z klockami Lego jest mylące. Liczby

algebraiczne wyglądają inaczej. Bardziej przypominają klocki z ruchomymi połączeniami, które mogą się sprzęgać w różny sposób. Jeśli rozłożymy taki klocek w jeden sposób, otrzymamy sprzęgnięte ze sobą części, których nie da się już bardziej rozdzielić. Jeśli zaś rozłożymy go inaczej, to również otrzymamy niepodzielne części, tyle tylko że będą one teraz inne.

Rozważmy dwa przykłady. W pierwszym wykorzystamy zwyczajne liczby całkowite. Łatwo taki przykład zrozumieć, ale ma on pewne specyficzne cechy, których nie znajdziemy w przypadku innych liczb. Potem rozważymy przykład z prawdziwymi liczbami algebraicznymi.

Załóżmy, że żyjemy we wszechświecie, w którym istnieją wyłącznie liczby z ciągu 1, 5, 9, 13, 17, 21, 25 i tak dalej – to znaczy liczby, które w naszym prawdziwym Wszechświecie mają postać $4k + 1$. Jeśli pomnożymy przez siebie dwie takie liczby, to uzyskamy liczbę tego samego rodzaju. Przyjmijmy, że za liczbę „pierwszą" będziemy uważali taką, która nie jest iloczynem dwóch mniejszych liczb *tego rodzaju*. Na przykład 25 nie jest liczbą pierwszą, ponieważ jest równe 5×5, a 5 jest liczbą z naszej listy. Natomiast 21 *jest* liczbą pierwszą w takim nowym znaczeniu, ponieważ czynniki, na które się zwykle rozkłada, czyli liczby 3 i 7, nie znajdują się na liście. Czynniki te są liczbami postaci $4k + 3$, a nie $4k + 1$. Nietrudno zauważyć, że każda liczba z naszej listy musi być iloczynem liczb pierwszych rozumianych zgodnie z nową definicją. Przyczyną tego jest fakt, że czynniki danej liczby, jeśli istnieją, muszą być coraz mniejsze. W końcu proces podziału na czynniki dobiegnie końca, a gdy to się stanie, otrzymane czynniki będą liczbami pierwszymi.

Jednak taki rozkład na czynniki pierwsze nie jest jednoznaczny. Weźmy na przykład liczbę 4389, która jest równa $4 \times 1097 + 1$, a więc ma odpowiednią postać. Istnieją trzy różne rozkłady na czynniki będące liczbami o wymaganej postaci:

$$4389 = 21 \times 209 = 33 \times 133 = 57 \times 77.$$

Okazuje się, że wszystkie te czynniki są liczbami pierwszymi w rozumieniu zgodnym z naszą bieżącą definicją. Na przykład 57 jest

liczbą pierwszą, ponieważ jej normalne czynniki 3 i 19 nie są liczbami wymaganej postaci. To samo dotyczy liczb 21, 33, 77, 133 i 209. Wyjaśnijmy zatem, skąd wziął się ten brak jednoznaczności. W normalnym zbiorze liczb całkowitych mamy:

$$4389 = 3 \times 7 \times 11 \times 19,$$

ale wszystkie te czynniki mają „niewłaściwą" postać $4k + 3$. Trzy różne rozkłady na czynniki pierwsze – zgodnie z nową definicją – powstają w wyniku połączenia tych liczb w pary:

$$(3 \times 7) \times (11 \times 19),$$

$$(3 \times 11) \times (7 \times 19),$$

$$(3 \times 19) \times (7 \times 11).$$

Musimy się posłużyć parami, ponieważ iloczyn dwóch liczb postaci $4k + 3$ jest liczbą postaci $4k + 1$.

Przykład ten pokazuje, że rozumowanie, iż „rozkład na czynniki musi być jednoznaczny, ponieważ są one najmniejszymi elementami", jest błędne. To prawda, że gdzieś tam istnieją *mniejsze* elementy (na przykład $21 = 3 \times 7$), ale nie należą one do naszego zbioru liczb. Główny powód, dla którego przykład ten nie jest reprezentatywny, stanowi to, że chociaż mnożenie dwóch liczb postaci $4k + 1$ daje liczbę takiej samej postaci, to nie jest to już prawdą w przypadku dodawania. Na przykład $5 + 5 = 10$, a 10 nie jest liczbą z naszej listy. Używając języka algebry, powiedzielibyśmy, że nasz zbiór liczb nie jest pierścieniem.

Drugi przykład nie ma tej wady, ale jest za to nieco trudniejszy do przeanalizowania. Rozważmy pierścień liczb algebraicznych całkowitych wielomianu $x^2 - 15$. Pierścień ten tworzą wszystkie liczby postaci $a + b\sqrt{15}$, gdzie a i b są liczbami całkowitymi. W takim pierścieniu liczbę 10 da się rozłożyć na dwa sposoby:

$$10 = 2 \times 5 = \left(5 + \sqrt{15}\right) \times \left(5 - \sqrt{15}\right)$$

Można udowodnić, że wszystkie te czynniki: 2, 5, $5 + \sqrt{15}$ oraz $5 - \sqrt{15}$ są liczbami pierwszymi[56].

Wszystko to rozumiemy obecnie dużo lepiej niż w 1847 roku, ale Liouville wcale nie musiał długo czekać, by przekonać się, iż jego wątpliwości były uzasadnione. Już dwa tygodnie później Pierre Wantzel poinformował akademię, że jednoznaczność rozkładu obowiązuje dla niektórych małych wartości *p*, ale jego metoda dowodzenia zawiodła dla potęgi 23. Niedługo potem Liouville oznajmił, że rozkład na czynniki pierwsze *nie jest* jednoznaczny w przypadku liczb cyklotomicznych całkowitych odpowiadających *p* = 23. Ernst Kummer odkrył to już trzy lata wcześniej, ale nikomu nic nie powiedział, ponieważ próbował znaleźć metodę obejścia tej przeszkody. Dowód Lamégo okazał się poprawny dla małych wartości *p*, w tym dla kilku nowych: 11, 13, 17, 19, ale ogólnie był błędny. Była to lekcja poglądowa pokazująca, że nie należy zakładać, iż wiarygodnie wyglądające stwierdzenia matematyczne są oczywiste. Czasami może się okazać, że wcale nie są prawdziwe.

Kummer rozmyślał nad wielkim twierdzeniem Fermata, podążając w podobnym kierunku co Lamé. Zauważył potencjalną przeszkodę, przyjrzał się jej uważnie i stwierdził, że uniemożliwia ona dalsze prace w tym kierunku. Znalazł konkretny przykład niejednoznacznego rozkładu na czynniki pierwsze liczb cyklotomicznych całkowitych opartych na pierwiastkach 23 stopnia z jedynki. Kummer tak łatwo się jednak nie poddał i postanowił znaleźć obejście tej przeszkody – a przynajmniej jakiś sposób na uniknięcie jej najgorszych cech. Jego pomysł można szczególnie dobrze wytłumaczyć na naszym przykładzie liczb postaci $4k + 1$. Chodzi o to, że w celu przywrócenia jednoznaczności rozkładu musimy do rozważanego zbioru dodać pewne *nowe* liczby. W naszym przykładzie potrzebne nam są brakujące liczby postaci $4k + 3$. Możemy też pójść na całość i dorzucić za jednym zamachem także

parzyste liczby całkowite. Otrzymamy wtedy zbiór liczb całkowitych zamknięty ze względu na dodawanie i mnożenie. Innymi słowy, jeśli dodamy do siebie dwie liczby całkowite lub je przez siebie pomnożymy, to uzyskany wynik będzie liczbą całkowitą.

Kummer wymyślił pewną odmianę tej idei. Możemy na przykład odzyskać jednoznaczność rozkładu na czynniki pierwsze w pierścieniu wszystkich liczb postaci $a+b\sqrt{15}$, dodając do niego nową liczbę: $\sqrt{5}$. Okazuje się, że jeśli chcemy mieć pierścień, to musimy jeszcze dorzucić $\sqrt{3}$. Teraz

$$2=\left(\sqrt{5}+\sqrt{3}\right)\times\left(\sqrt{5}-\sqrt{3}\right)$$

$$5=\sqrt{5}\times\sqrt{5}$$

oraz

$$5+\sqrt{15}=\sqrt{5}\times\left(\sqrt{5}+\sqrt{3}\right)$$

$$5-\sqrt{15}=\sqrt{5}\times\left(\sqrt{5}-\sqrt{3}\right)$$

Zatem oba rozkłady na czynniki powstają w wyniku pogrupowania na dwa różne sposoby czterech liczb $\sqrt{5}$, $\sqrt{5}$, $\sqrt{5}+\sqrt{3}$ i $\sqrt{5}-\sqrt{3}$.

Kummer nazwał te nowe elementy liczbami idealnymi, ponieważ w jego ogólnym sformułowaniu nie były one w pełni liczbami. Były symbolami, które zachowywały się jak liczby. Dowiódł, że każdą liczbę cyklotomiczną całkowitą można rozłożyć jednoznacznie na czynniki będące idealnymi liczbami pierwszymi. Jest to dość specyficzny układ: ani liczby cyklotomiczne całkowite, ani liczby idealne nie mają jednoznacznego rozkładu na czynniki pierwsze. Jeżeli jednak użyjemy liczb idealnych jako składników rozkładu na

czynniki pierwsze liczb cyklotomicznych całkowitych, to taki rozkład będzie jednoznaczny.

Później Richard Dedekind przedstawił bardziej elegancką wersję procedury Kummera i właśnie takie podejście stosujemy obecnie. Z każdą liczbą idealną spoza rozważanego pierścienia Dedekind powiązał *zbiór* liczb należących do pierścienia. Nazwał ten zbiór ideałem. Każda liczba w pierścieniu definiuje pewien ideał: składa się on z wszystkich jej wielokrotności. Jeśli rozkład na czynniki pierwsze jest jednoznaczny, to wszystkie ideały mają taką właśnie postać. Jeśli natomiast nie jest jednoznaczny, to istnieją ideały dodatkowe. Możemy wyznaczyć iloczyny i sumy ideałów oraz ideałów pierwszych, a Dedekind udowodnił, że rozkład na czynniki pierwsze *ideałów* jest jednoznaczny dla wszystkich pierścieni liczb algebraicznych całkowitych. Wynika z tego, że w większości wypadków powinniśmy działać na ideałach, a nie na samych liczbach algebraicznych. To oczywiście wprowadza nowe komplikacje, ale jedyną alternatywą jest brak dalszych postępów.

Kummer zdołał rozwinąć swoją koncepcję liczb idealnych na tyle, że udało mu się udowodnić pewną wersję wielkiego twierdzenia Fermata rozszerzoną o dodatkowe hipotezy. Jednak pozostali śmiertelnicy uważali, że liczby idealne są zbyt trudne, a może nawet odrobinę mistyczne. Dopiero prace Dedekinda spowodowały, że liczby te stały się zrozumiałe i algebraiczna teoria liczb ruszyła do przodu. Jedną z ważnych idei, które powstały w wyniku tych prac, była możliwość pomiaru tego, jak mało jednoznaczny jest podział na czynniki w pierścieniu liczb algebraicznych całkowitych. Każdemu takiemu pierścieniowi odpowiada liczba naturalna nazywana *rzędem grupy klas ideałów*. Jeśli liczba ta jest równa 1, to rozkład na czynniki pierwsze jest jednoznaczny, a w przeciwnym razie – nie. Im większa wartość rzędu grupy klas ideałów, tym „mniej jednoznaczny" jest rozkład na czynniki pierwsze.

Możliwość ilościowego określenia niejednoznaczności rozkładu na czynniki pierwsze była olbrzymim krokiem naprzód i wystarczyło włożyć w te prace jeszcze trochę wysiłku, żeby uratować strategię Lamégo – przynajmniej w niektórych przypadkach. W 1850 roku Kummer ogłosił, że udało mu się udowodnić wielkie twierdzenie

Fermata dla wielu liczb pierwszych, a mianowicie dla tych, które nazwał regularnymi. Wśród liczb pierwszych mniejszych od 100 jedynie 37, 59 i 67 są nieregularne. Dla wszystkich pozostałych liczb pierwszych z tego przedziału i wielu z tych, które znajdują się poza nim, jego metody pozwalają udowodnić wielkie twierdzenie Fermata. Definicja regularnej liczby pierwszej bazuje na pojęciu rzędu grupy klas ideałów: liczba pierwsza jest regularna, jeśli nie jest dzielnikiem rzędu grupy klas ideałów odpowiedniego pierścienia cyklotomicznych liczb całkowitych. Wynika więc z tego, że w przypadku regularnej liczby pierwszej rozkład na czynniki pierwsze nie jest jednoznaczny, ale ów brak jednoznaczności nie dotyczy w żaden istotny sposób tej liczby.

Kummer stwierdził, że istnieje nieskończenie wiele regularnych liczb pierwszych, ale hipoteza ta pozostaje do dziś nieudowodniona. Jak na ironię, w 1915 roku K.L. Jensen udowodnił, że istnieje nieskończenie wiele nieregularnych liczb pierwszych. Z badań z zakresu analizy matematycznej wyłoniło się dość dziwne kryterium regularności liczb pierwszych. Ma ono związek z ciągiem liczbowym odkrytym niezależnie przez japońskiego matematyka Seki Takakazu (znanego też jako Seki Kōwa) i szwajcarskiego matematyka Jakoba Bernoulliego. Liczby tego ciągu nazywamy obecnie liczbami Bernoulliego. Z tego kryterium wynika, że pierwsze dziesięć nieregularnych liczb pierwszych to: 37, 59, 67, 101, 103, 131, 149, 157, 233 i 257. Badając głębiej strukturę cyklotomicznych liczb całkowitych, Dmitrij Mirimanow w 1893 roku zdołał udowodnić wielkie twierdzenie Fermata dla pierwszej nieregularnej liczby pierwszej, czyli dla 37. Do roku 1905 dowiódł prawdziwości tego twierdzenia dla liczb do wartości $p = 257$. Harry Vandiver opracował algorytmy komputerowe pozwalające dojść jeszcze dalej. Wykorzystując te metody, John Selfridge i Bary Pollack w 1967 roku udowodnili twierdzenie Fermata aż do potęgi 25 000, a w 1976 roku Samuel S. Wagstaff zdołał przesunąć tę granicę aż do wartości 100 000.

Matematycy gromadzili coraz więcej dowodów przemawiających za poprawnością wielkiego twierdzenia Fermata, ale wynikało z nich głównie to, że gdyby twierdzenie to okazało się jednak fałszywe, to

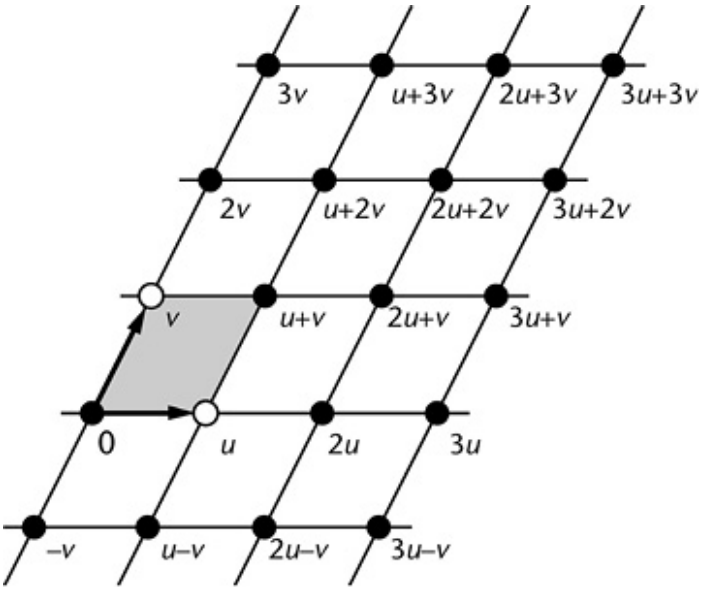
odpowiedni kontrprzykład – przykład pokazujący, że nie jest ono spełnione – musiałby być tak olbrzymi, że raczej nikomu nie uda się go znaleźć. Inny wniosek płynący z tych prac był taki, że metody podobne do podejścia Kummera natrafiają na te same problemy, z jakimi borykali się pierwsi matematycy próbujący udowodnić twierdzenie Fermata: większe potęgi wymagają specjalnego, skomplikowanego podejścia. W rezultacie prace na tym froncie badań po pewnym czasie zamarły.

Gdy prace nad rozwiązaniem jakiegoś problemu matematycznego stają w martwym punkcie, warto posłuchać rady Poincarégo: zostaw to i zajmij się czymś innym. Przy odrobinie szczęścia i sprzyjających wiatrach w końcu pojawi się jakiś nowy pomysł. Specjaliści od teorii liczb nie podjęli takiej decyzji świadomie, niemniej jednak tak właśnie postąpili. I tak jak przewidział Poincaré, przyniosło to owoce. Niektórzy teoretycy liczb zainteresowali się krzywymi eliptycznymi, o których mówiliśmy w rozdziale 6. Co ciekawe, ostatecznie okazało się, że ten obszar badań w zaskakujący sposób wiąże się z wielkim twierdzeniem Fermata, czego ukoronowaniem był dowód Wilesa. Aby opisać ten związek, musimy wprowadzić jeszcze jedno pojęcie, a mianowicie funkcję modularną. W związku z tym będziemy teraz musieli zagłębić się trochę w szczegóły matematyczne, ale za tymi ideami kryje się ciekawa opowieść, a nam wystarczy omówienie tych koncepcji na poziomie ogólnym. Proszę jeszcze o odrobinę cierpliwości.

W rozdziale 6 dowiedzieliśmy się, że teoria funkcji eliptycznych wywarła głęboki wpływ na analizę zespoloną. W latach trzydziestych XIX wieku Joseph Liouville odkrył, że różnorodność funkcji eliptycznych jest dość ograniczona. Dla dowolnych dwóch okresów istnieje specjalna funkcja eliptyczna, zwana funkcją Weierstrassa, która charakteryzuje się tym, że każda inna funkcja eliptyczna o takich samych okresach jest jej prostą odmianą. Wynika z tego, że jedynymi funkcjami o podwójnym okresie, które trzeba zrozumieć, są funkcje Weierstrassa – po jednej dla każdej pary okresów.

Podwójnie okresową strukturę funkcji eliptycznej można przedstawić geometrycznie za pomocą sieci na płaszczyźnie zespolonej. Sieć taką wyznaczają wszystkie kombinacje obydwu okresów *u* i *v* postaci *mu* + *nv*, gdzie *m* i *n* są liczbami całkowitymi (zob. ryc. 30). Jeśli wybierzemy dowolną liczbę zespoloną *z* i dodamy do niej jeden z punktów takiej sieci, to wartość funkcji eliptycznej w nowym punkcie będzie taka sama jak w wyjściowym punkcie *z*. Innymi słowy, funkcja eliptyczna ma taką samą symetrię jak jej sieć.

Specjaliści od analityki matematycznej odkryli znacznie bogatsze źródło symetrii na płaszczyźnie zespolonej: przekształcenia Möbiusa. Zmieniają one punkt *z* na (*az* + *b*)/(*cz* + *d*), gdzie *a*, *b, c* i *d* są stałymi liczbami zespolonymi. Opisane przed chwilą symetrie sieci są szczególnymi przypadkami przekształceń Möbiusa, ale istnieją także inne ich rodzaje. W takim bardziej ogólnym przypadku w dalszym ciągu istnieją jednak zbiory punktów analogiczne do naszej sieci. Sieć dzieli płaszczyznę euklidesową na jednakowe obszary – w tym przykładzie są to równoległoboki, których rogi znajdują się w punktach sieci (zob. ryc. 26 i 30). Wykorzystując przekształcenia Möbiusa, możemy podzielić odpowiednią powierzchnię nieeuklidesową, czyli płaszczyznę hiperboliczną, na jednakowe obszary. Geometrię tę można przedstawić za pomocą obszaru płaszczyzny zespolonej, na którym linie proste przyjmują postać łuków okręgów.



Ryc. 30. Sieć utworzona ze wszystkich kombinacji całkowitych obydwu okresów funkcji eliptycznej

W geometrii hiperbolicznej możliwe są wysoce symetryczne podziały na jednakowe elementy. Dla każdego z nich możemy skonstruować funkcje zespolone powtarzające te same wartości na obszarze każdego elementu. Funkcje te, będące naturalnym uogólnieniem funkcji eliptycznych, nazywamy funkcjami modularnymi. Geometria hiperboliczna to bardzo bogata dziedzina i istnieje w niej dużo więcej możliwości podziału płaszczyzny na jednakowe obszary niż w geometrii euklidesowej. Specjaliści od analizy zespolonej zaczęli więc intensywnie badać geometrię nieeuklidesową. W wyniku ich prac ujawnił się głęboki związek między analizą matematyczną i teorią liczb. Okazało się, że funkcje modularne mają takie samo znaczenie dla funkcji eliptycznych, jak funkcje trygonometryczne dla okręgu.

Jak pamiętamy, okrąg jednostkowy składa się z punktów (*x*, *y*) takich, że $x^2 + y^2 = 1$. Załóżmy teraz, że *A* jest liczbą rzeczywistą, i przyjmijmy, że

$$x = \cos A$$

oraz

$$y = \sin A$$

Z definicji funkcji sinus i cosinus wynika, że punkt ten leży na okręgu jednostkowym. Co więcej, w takiej postaci możemy zapisać każdy punkt na tym okręgu. Mówiąc językiem matematyków, powiedzielibyśmy, że funkcje trygonometryczne *parametryzują* okrąg. Coś bardzo podobnego obserwujemy dla funkcji modularnych. Jeśli zdefiniujemy *x* i *y* za pomocą odpowiednich funkcji modularnych parametru *A*, to tak określony punkt będzie leżał na krzywej eliptycznej – zawsze takiej samej, bez względu na to, jaką wartość przyjmie parametr *A*. Istnieją bardziej abstrakcyjne sposoby pozwalające wyrazić to precyzyjniej i specjaliści właśnie nimi się posługują, bo tak jest wygodniej, ale takie ujęcie pozwala dostrzec podobieństwo z trygonometrią i okręgiem. Ten związek przypisuje każdej funkcji modularnej pewną krzywą eliptyczną. Zauważmy jednak, że różnych funkcji modularnych jest bardzo dużo – są nimi wszystkie symetryczne podziały płaszczyzny hiperbolicznej na jednakowe elementy. Zatem z funkcjami modularnymi da się powiązać bardzo dużo krzywych eliptycznych. Które krzywe eliptyczne można w ten sposób uzyskać? Okazało się, że jest to kluczowe pytanie.

To „brakujące ogniwo” zauważono po raz pierwszy w 1975 roku, gdy Yves Hellegouarch zwrócił uwagę na dziwny związek między wielkim twierdzeniem Fermata i krzywymi eliptycznymi. Gerhard Frey poszedł tym tropem dalej i wyniki swoich prac przedstawił w artykułach opublikowanych w 1982 i 1986 roku. Przyjmijmy jak zwykle, że *p* jest nieparzystą liczbą pierwszą, i załóżmy – mając nadzieję doprowadzić całą rzecz do sprzeczności – że istnieją niezerowe liczby całkowite *a*, *b* i *c* spełniające równanie Fermata, a więc takie, że $a^p + b^p = c^p$. Teraz pora wyciągnąć teatralnym gestem królika z kapelusza: rozważmy krzywą eliptyczną

$$y^2 = x\left(x - a^p\right)\left(x - b^p\right)$$

Jest to tak zwana krzywa eliptyczna Freya. Frey zastosował do niej wszystkie znane mu techniki i z jego badań wyłonił się ciąg coraz bardziej zdumiewających zbiegów okoliczności. Jego hipotetyczna krzywa eliptyczna faktycznie jest bardzo dziwna. Wydaje się bezsensowna. Frey udowodnił, że jest tak bardzo pozbawiona sensu, iż nie może istnieć. Z tego wynika oczywiście poprawność wielkiego twierdzenia Fermata, ponieważ doszliśmy do sprzeczności, tak jak planowaliśmy.

W tym wszystkim jest jednak pewna luka, o której istnieniu Frey doskonale wiedział. Aby dowieść, że ta hipotetyczna krzywa eliptyczna nie istnieje, musimy pokazać, że gdyby istniała, to byłaby modularna – to znaczy, że byłaby jedną z krzywych wynikających z funkcji modularnych. Jak już wiemy, takich krzywych jest bardzo dużo i w tamtym czasie nikomu nie udało się jeszcze znaleźć krzywej eliptycznej, która *nie byłaby* modularna. Wszystko wskazywało więc na to, że krzywa Freya również powinna być modularna – była to jednak krzywa hipotetyczna, liczby *a*, *b* i *c* nie były znane, a na domiar złego, gdyby faktycznie *była* krzywą modularną, to w ogóle by nie istniała. Ze wszystkimi tymi problemami można się było rozprawić tylko w jeden sposób – należało udowodnić, że *każda* krzywa eliptyczna jest modularna. W takim razie krzywa Freya, nieważne że hipotetyczna, musiałaby być modularna, jeśli istnieje. A jeśli nie istnieje, to dowód również zostałby dokończony.

Stwierdzenie, że każda krzywa eliptyczna jest modularna, jest znane jako hipoteza Taniyamy–Shimury. Nazwano ją tak na cześć dwóch japońskich matematyków Yutaki Taniyamy i Goro Shimury. Uczeni ci spotkali się zupełnie przypadkiem w bibliotece, gdy chcieli wypożyczyć tę samą książkę z takiego samego powodu. Tak rozpoczęła się ich długa współpraca. W 1955 roku Taniyama uczestniczył w konferencji matematycznej zorganizowanej w Tokio, na której poproszono młodszych uczestników o przygotowanie listy otwartych pytań. Taniyama zgłosił cztery problemy, wszystkie dotyczące związku między funkcjami modularnymi i krzywymi

eliptycznymi. Jeszcze przed konferencją wyznaczył kilka wartości związanych z pewną określoną funkcją modularną i zauważył, że te same liczby pojawiają się w związku z pewną określoną krzywą eliptyczną. Tego rodzaju zbiegi okoliczności świadczą często o tym, że wcale nie mamy do czynienia z przypadkiem i musi istnieć jakieś sensowne ich wytłumaczenie. Obecnie wiemy, że równość tych liczb oznacza, iż taka krzywa eliptyczna jest modularna. Prawdę mówiąc, właśnie tak się je obecnie najczęściej definiuje w literaturze fachowej. Tak czy inaczej, Taniyamę tak to zaintrygowało, że wyznaczył te wartości jeszcze dla kilku innych funkcji modularnych i również w ich przypadku przekonał się, że odpowiadają one określonym krzywym eliptycznym.

Zaczął się zastanawiać, czy podobna prawidłowość obowiązuje dla wszystkich funkcji eliptycznych. Większość specjalistów z tej dziedziny uważała, że byłoby to zbyt piękne, by mogło być prawdziwe – że jest to jedynie marzenie niepoparte dowodami. Shimura był jednym z nielicznych, którzy sądzili, że ta hipoteza zasługuje na uwagę. Jednak w latach 1957–1958 wyjechał na rok do Princeton i podczas jego nieobecności Taniyama popełnił samobójstwo. Pozostawił list, w którym między innymi napisał: „Jeśli chodzi o powód mojego samobójstwa, to sam nie całkiem go rozumiem, ale nie jest ono wynikiem żadnego szczególnego wydarzenia ani określonej sprawy. Mogę jedynie powiedzieć, że towarzyszy mi nastrój utraty wiary w moją przyszłość”[57]. W tym okresie zamierzał się ożenić i niedoszła panna młoda, Misako Suzuki, zabiła się mniej więcej miesiąc później. W jej liście pożegnalnym znalazło się takie zdanie: „Teraz, gdy odszedł, ja też muszę odejść, żeby do niego dołączyć”[58].

Shimura kontynuował pracę nad hipotezą i w miarę jak gromadził kolejne potwierdzające ją dowody, ogarniało go coraz silniejsze przekonanie, że może jednak być prawdziwa. Większość specjalistów w tej dziedzinie nie podzielała takiej opinii. Simon Singh przytacza wywiad z Shimurą, w którym uczony wspomina, jak próbował wyjaśnić to jednemu z kolegów:

– Słyszałem, że podobno pan uważa, że niektóre równania eliptyczne wiążą się z formami modułowymi – spytał profesor.
– Nie, nie zrozumiał pan – odparł Shimura. – Nie chodzi tylko o niektóre równania eliptyczne, ale o wszystkie równania eliptyczne![59]

Mimo dużej nieufności środowiska Shimura trwał przy swoim i po wielu latach hipoteza stała się na tyle znana, że zaczęto ją nazywać hipotezą Taniyamy–Shimury. Wówczas André Weil, jeden z wielkich teoretyków liczb XX stulecia, znalazł kolejne dowody na potwierdzenie prawdziwości tej hipotezy i opublikował wyniki swoich badań, wyrażając przekonanie, że hipoteza Taniyamy–Shimury może być prawdziwa. Od tej chwili matematycy zaczęli ją nazywać hipotezą Taniyamy–Shimury–Weila. W takim brzmieniu nazwa nigdy się jednak nie przyjęła i w kontekście tej hipotezy nazwiska trójki uczonych pojawiają się w różnej kolejności i różnych konfiguracjach. Pozostańmy więc przy nazwie „hipoteza Taniyamy–Shimury".

W latach sześćdziesiątych inny gigant matematyki, Robert Langlands, uświadomił sobie, że hipotezę Taniyamy–Shimury można uważać za jeden z elementów znacznie szerszego i ambitniejszego programu połączenia ze sobą algebraicznej i analitycznej teorii liczb. Sformułował mnóstwo hipotez związanych z tą koncepcją, które obecnie znane są jako program Langlandsa. Był on jeszcze mniej pewny od hipotezy Taniyamy–Shimury, ale miał pewien urok i elegancję – wydawał się jednym z owych rozwiązań matematycznych, które powinny być prawdziwe, ponieważ są tak piękne. W latach siedemdziesiątych świat matematyki przyzwyczaił się do piękna programu Langlandsa i zaczęto go uważać za jeden z podstawowych celów algebraicznej teorii liczb. Wydawało się, że program ten wytycza właściwy kierunek dalszych badań – brakowało tylko kogoś, kto uczyni pierwszy krok.

Właśnie wtedy Frey zauważył, że zastosowanie hipotezy Taniyamy–Shimury do jego krzywej eliptycznej pozwoli udowodnić wielkie twierdzenie Fermata. Tymczasem zdążył się jednak pojawić kolejny problem związany z jego koncepcją. Gdy w 1984 roku wygłaszał

wykład poświęcony temu zagadnieniu, jeden ze słuchaczy zauważył lukę w kluczowej argumentacji: otóż krzywa ta jest tak dziwaczna, że nie może być modularna. Jean-Pierre Serre, jeden z czołowych badaczy w tej dziedzinie, szybko wypełnił tę lukę, ale musiał się przy tym powołać na inną nieudowodnioną jeszcze hipotezę – na hipotezę o redukcji poziomu. W 1986 roku Ken Ribet zdołał jednak w końcu udowodnić hipotezę o redukcji poziomu i ostatnią przeszkodą stojącą na drodze do uzyskania dowodu wielkiego twierdzenia Fermata była hipoteza Taniyamy–Shimury. Nagle opinia środowiska matematycznego na temat tej hipotezy zaczęła się zmieniać. Serre wysunął przypuszczenie, że wielkie twierdzenie Fermata zostanie udowodnione w ciągu najbliższych dziesięciu lat. Jak do tego miałoby dojść, to zupełnie inna kwestia, ale w powietrzu wisiało coś, co napawało nadzieją. Metody związane z funkcjami modularnymi stały się tak potężne, że komuś musiało się w końcu poszczęścić w pracach nad dokończeniem dzieła Freya.

Tym kimś okazał się Andrew Wiles. W programie telewizyjnym poświęconym dowodowi wielkiego twierdzenia Fermata tak wspominał swoją młodość:

> Miałem dziesięć lat, gdy […] natrafiłem na książkę poświęconą matematyce, w której opisano historię tego problemu [wielkiego twierdzenia Fermata]. Dowiedziałem się, że ktoś wysunął to twierdzenie trzysta lat temu, ale nikt nie przedstawił jego dowodu i nie wiadomo, czy w ogóle dowód taki istnieje, mimo że wszyscy od tego czasu go szukają. I oto miałem przed sobą problem, który ja – dziesięciolatek – mogłem zrozumieć, a jednak żadnemu z wielkich matematyków przeszłości nie udało się go rozwiązać. Oczywiście od tej chwili próbowałem sam go rozwiązać. Było to duże wyzwanie i piękny problem.

W 1971 roku Wiles ukończył studia matematyczne na Uniwersytecie Oksfordzkim i przeniósł się na uniwersytet

w Cambridge, by zrobić tam doktorat. Jego opiekun John Coates przekonał go (słusznie), że udowodnienie wielkiego twierdzenia Fermata jest zbyt trudnym tematem na pracę doktorską. Zamiast tego Wiles zajął się więc krzywymi eliptycznymi, które uważano wówczas za bardzo obiecujący obszar badań. W latach 1985–1986 przebywał w Paryżu w Institut des Hautes Études Scientifiques (Instytucie Zaawansowanych Badań Naukowych), jednym z czołowych ośrodków matematycznych świata. Większość wielkich uczonych trafia w to miejsce na jakimś etapie swojej kariery. Gdy jest się matematykiem, właśnie tam należy przebywać. Jednym z uczonych odwiedzających wówczas instytut był Ribet. Wilesa zafascynował przedstawiony przez Ribeta dowód hipotezy o redukcji poziomu. Dzięki niemu mógł dalej zajmować się szanowaną dziedziną krzywych eliptycznych, szukając dowodu hipotezy Taniyamy–Shimury, a jednocześnie zbliżać się do realizacji swojego marzenia z dzieciństwa o udowodnieniu wielkiego twierdzenia Fermata.

Ponieważ o istnieniu takiego związku wiedzieli wszyscy zajmujący się tym obszarem badań, istniały pewne powody do obaw. Przypuśćmy, że Wiles zdołałby przedstawić niemal pełny dowód, w którym byłoby tylko kilka mniej istotnych luk wymagających jeszcze trochę pracy. Przypuśćmy też, że ktoś by się o tym dowiedział i wypełnił te luki. Wtedy, formalnie rzecz biorąc, to właśnie ta osoba udowodniłaby wielkie twierdzenie Fermata. Matematycy najczęściej tak się nie zachowują, ale gdy nagroda jest tak wielka, warto się zabezpieczyć. Dlatego Wiles prowadził swoje badania w tajemnicy, co w świecie matematyków należy do rzadkości. Nie można powiedzieć, że nie ufał swoim kolegom. Nie chciał jedynie podejmować najmniejszego ryzyka, że tuż przed metą ktoś go wyprzedzi.

Zaszył się na poddaszu swojego domu, gdzie urządził sobie gabinet, i pracował tam samotnie przez siedem lat. Tylko jego żona i szef wydziału wiedzieli, nad czym pracuje. W spokoju i odosobnieniu atakował ten problem każdą metodą, jaka wpadła mu w ręce, aż w końcu mury zamku zaczęły drżeć pod intensywnym ostrzałem. W 1991 roku Coates zwrócił mu uwagę na pewne wyniki,

które uzyskał Matthias Flach. Pod naporem ataku szczelina w murze zaczęła się poszerzać coraz szybciej.

W 1993 roku dowód był gotowy. Teraz trzeba było go pokazać światu. Wiles chciał zachować nadal ostrożność i nie narażać się na ryzyko, że przedstawi swoje rozwiązanie, a zaraz potem wypłynie jakiś błąd – coś takiego przytrafiło się Yoichiemu Miyaoce w 1988 roku, gdy w prasie pojawiła się informacja, że ukończył dowód, a potem znaleziono w jego pracy błąd. Wiles postanowił więc, że wygłosi serię trzech wykładów w Instytucie Isaaca Newtona w Cambridge – nowo otwartym międzynarodowym centrum badań matematycznych. Tytuł wykładów brzmiał niewinnie i dość zawile: *Formy modularne, krzywe eliptyczne i teoria Galois*. Mało kto dał się jednak zwieść – wszyscy wiedzieli, że Wiles ma w zanadrzu coś wielkiego.

Na trzecim wykładzie Wiles nakreślił zarys dowodu szczególnego przypadku hipotezy Taniyamy–Shimury. Odkrył, że wystarczy udowodnić coś nieco mniej ambitnego – mianowicie pokazać, że krzywa Freya, jeśli istnieje, musi należeć do specjalnej klasy krzywych eliptycznych, tak zwanych krzywych półstabilnych, i udowodnić, że wszystkie krzywe *tej klasy* muszą być modularne. Następnie udowodnił obie te rzeczy. Pod koniec wykładu zapisał na tablicy wniosek, który wynika bezpośrednio z tego, co przed chwilą pokazał. Wnioskiem tym było wielkie twierdzenie Fermata.

Gdy Shimura dowiedział się, czego dokonał Wiles, skomentował to krótko i rzeczowo: „A nie mówiłem?”.

Gdybyż to tylko było takie proste. Los postanowił jednak spłatać Wilesowi figla. Teraz dowód musieli jeszcze sprawdzić recenzenci i jak to zwykle bywa, zażyczyli sobie wyjaśnienia kilku kwestii. Wiles poradził sobie z większością ich komentarzy, ale jedna z uwag zmusiła go do przemyślenia wszystkiego na nowo. Pod koniec 1993 roku wystosował oświadczenie, w którym stwierdził, że wycofuje swój dowód do czasu, gdy uda mu się wypełnić wykrytą lukę logiczną. Teraz musiał jednak pracować w pełnym blasku

reflektorów, a więc znalazł się w sytuacji, której tak bardzo chciał uniknąć.

W marcu 1994 roku wciąż nie było widać poprawionej wersji dowodu i Faltings wyraził powszechnie spotykaną wówczas wśród matematyków opinię: „Gdyby [naprawienie dowodu] było proste, już dawno przedstawiłby poprawioną wersję. Ściśle rzecz biorąc, w chwili jego ogłoszenia nie był to jeszcze dowód". Coś podobnego powiedział też Weil: „Wierzę, że wpadł na kilka dobrych pomysłów […] ale dowodu wciąż nie ma. […] Udowodnienie wielkiego twierdzenia Fermata jest jak wejście na Mount Everest. Jeśli ktoś stawia sobie za cel zdobycie najwyższego szczytu świata, ale musi się wycofać sto metrów przed szczytem, to jednak nie wspiął się na Everest". Wszyscy przeczuwali, jak to się skończy. Znali takie przypadki z przeszłości. Dowód się załamał i całą pracę trzeba będzie zacząć od początku, a wielkie twierdzenie Fermata nadal będzie czekało na zdobywcę.

Wiles nie chciał jednak uznać porażki i zaprosił do współpracy swojego byłego studenta Richarda Taylora. Razem udało im się dotrzeć do sedna problemu – okazało się, że w ich przypadku wyniki Flacha nie do końca nadają się do wykorzystania. Próbowali zmodyfikować metody Flacha, ale nic nie działało. I wtedy, w przypływie inspiracji, Wiles nagle zrozumiał, co im przeszkadza w osiągnięciu celu. „Pojąłem, że to, co uniemożliwia zastosowanie metody Flacha, wymaga użycia innego podejścia, które próbowałem wykorzystać już wcześniej". To tak jak gdyby żołnierze oblegający zamek uświadomili sobie nagle, że za pomocą tarana nigdy go nie zdobędą, ponieważ obrońcy ciskają w nich kamieniami, ale przecież te same kamienie można załadować do katapulty i rozbić nimi bramę broniącą dostępu do środka.

W kwietniu 1995 roku nowy dowód został ukończony i tym razem nie było w nim żadnych luk ani błędów. Uczeni szybko opublikowali swoje wyniki w postaci dwóch artykułów w niezwykle prestiżowym czasopiśmie „Annals of Mathematics". Wiles zdobył międzynarodową sławę, kilka ważnych nagród oraz tytuł szlachecki… po czym wrócił do badań i pracuje, tak jak wcześniej.

W dowodzie Wilesa najważniejsze wcale nie jest to, że teraz już wiemy, iż wielkie twierdzenie Fermata jest prawdziwe. Jak wspomnieliśmy wcześniej, od tego twierdzenia nic ważnego nie zależy. Gdyby ktoś znalazł trzy 100-cyfrowe liczby i 250-cyfrową liczbę pierwszą niespełniające twierdzenia Fermata, to udowodniono by, że nie jest ono prawdziwe, ale nie ucierpiałby na tym żaden ważny dział matematyki. Oczywiście tak dużych liczb nie udałoby się znaleźć w bezpośrednim ataku z wykorzystaniem komputera, trzeba byłoby się więc wykazać wyjątkowym sprytem, ale znalezienie przykładu przeczącego twierdzeniu nikogo by specjalnie nie zmartwiło.

Najcenniejsze w pracy Wilesa okazało się udowodnienie półstabilnego przypadku hipotezy Taniyamy–Shimury. W ciągu następnych sześciu lat Christophe Breuil, Brian Conrad, Fred Diamond i Richard Taylor rozszerzyli metody Wilesa do tego stopnia, że można je obecnie stosować nie tylko do krzywych półstabilnych, ale do wszystkich krzywych eliptycznych. Uczeni ci udowodnili pełną hipotezę Taniyamy–Shimury, co całkowicie odmieniło teorię liczb. Od tej chwili, gdy ktoś natrafia na krzywą eliptyczną, wie ze stuprocentową pewnością, że jest to krzywa modularna, i może wykorzystać cały arsenał metod analitycznych. Metody te pozwoliły już rozwiązać wiele innych problemów z dziedziny teorii liczb i w przyszłości na pewno uda się dzięki nim jeszcze wiele dokonać.

54 Zob.: http://code.google.com/p/flyspeck/.

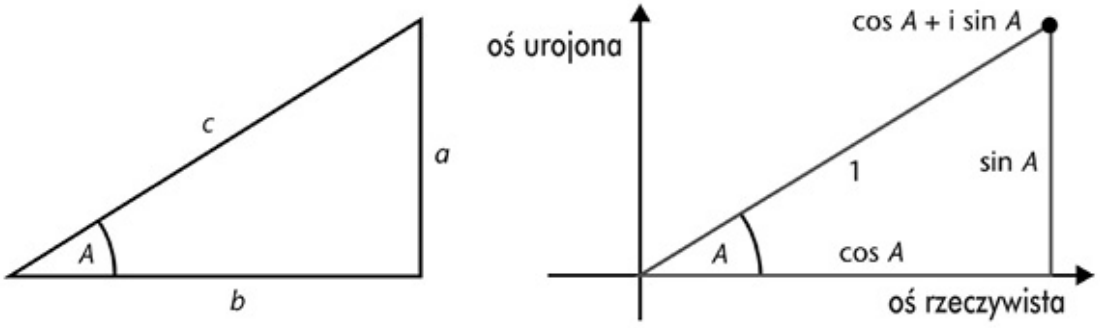
55 Jednym z pierwiastków *p*-tego stopnia z jedynki jest liczba zespolona:

$$\zeta = \cos 2\pi/p + i \sin 2\pi/p$$

natomiast pozostałe są jej kolejnymi potęgami: ζ^2, ζ^3, …, ζ^{p-1}. Aby zrozumieć, dlaczego tak jest, przypomnijmy sobie, że funkcje trygonometryczne sinus i cosinus definiuje się z wykorzystaniem trójkąta prostokątnego (zob. ryc. 56, po lewej). Oznaczając boki trójkąta tradycyjnymi symbolami *a*, *b* i *c*, możemy zdefiniować sinus (sin) i cosinus (cos) kąta *A* jako:

$$\sin A = a/c$$

$$\cos A = b/c$$



Ryc. 56. Po lewej: Definicja funkcji sinus i cosinus. Po prawej: Interpretacja na płaszczyźnie zespolonej.

Jeśli przyjmiemy, że *c* = 1 i umieścimy nasz trójkąt na płaszczyźnie zespolonej, tak jak na rycinie 56 po prawej stronie, to wierzchołek, w którym spotykają się boki *c* i *a*, jest liczbą zespoloną postaci:

$$\cos A + i \sin A$$

Nietrudno w tej sytuacji udowodnić, że dla dowolnych kątów *A* i *B* mamy:

$$\left(\cos A + i \sin A\right)\left(\cos B + i \sin B\right) = \cos\left(A + B\right) + i \sin\left(A + B\right)$$

a z tego otrzymujemy bezpośrednio wzór De Moivre'a:

$$\left(\cos A + i \sin A\right)^n = \cos nA + i \sin nA$$

obowiązujący dla dowolnej liczby naturalnej *n*. Zatem

$$\zeta^p = \left(\cos 2\pi/p + i \sin 2\pi/p\right)^p = \cos 2\pi + i \sin 2\pi = 1$$

a więc każda potęga 1, ζ, ζ^2, ζ^3, …, ζ^{p-1} jest *p*-tym pierwiastkiem z jedynki. Rozważamy tylko te liczby, ponieważ $\zeta^p = 1$ i uwzględniając wyższe potęgi, nie uzyskamy nowych wartości.

52 Wprowadźmy normę $\mathrm{N}\left(a + b\sqrt{15}\right) = a^2 - 15b^2$,

która ma tę wspaniałą cechę, że:

$$\mathrm{N}(xy) = \mathrm{N}(x)\mathrm{N}(y)$$

Wtedy

$$\mathrm{N}(2) = 4$$

$$\mathrm{N}(5) = 25$$

$$\mathrm{N}\left(5 + \sqrt{15}\right) = 10$$

$$\mathrm{N}\left(5 - \sqrt{15}\right) = 10$$

Każdy dzielnik właściwy dowolnej z tych czterech liczb musi mieć normę o wartości 2 albo 5 (są to dzielniki właściwe ich norm). Jednak równania $a^2 - 15^2 = 2$ i $a^2 - 15b^2 = 5$ nie mają rozwiązań w zbiorze liczb całkowitych. Zatem liczby te nie mają dzielników właściwych.

56 Dowód tego można znaleźć w wielu podręcznikach teorii liczb, na przykład w: Gareth A. Jones, J. Mary Jones, *Elementary Number Theory* (*Podstawy teorii liczb*), Springer, 1998, s. 227. W Internecie dowód ten przedstawiono na stronie: http://en.wikipedia.org/wiki/Infinite_descent#Non-solvability_of_r2_.2B_s4_.3D_t4.

57 Simon Singh, *Tajemnica Fermata: w poszukiwaniu rozwiązania najsłynniejszego matematycznego problemu świata*, przeł. Paweł Strzelecki, Prószyński i S-ka, Warszawa 1999, s. 145 (przyp. tłum.).

58 *Ibid.*, s. 145 (przyp. tłum.).

59 *Ibid.*, s. 146 (przyp. tłum.).

8. Chaos orbitalny

Problem trzech ciał

Z pewnego starego dowcipu można się dowiedzieć, że o stopniu zaawansowania teorii fizycznej świadczy to, przy jakiej liczbie oddziałujących ze sobą ciał teoria ta się załamuje. Newtonowskie prawo powszechnego ciążenia zaczyna mieć kłopoty przy trzech ciałach. Ogólna teoria względności nie radzi sobie z dwoma. Teoria kwantowa wspina się na szczyty swoich możliwości przy jednym ciele, a kwantowa teoria pola wpada w tarapaty nawet wtedy, gdy nie ma żadnych ciał – gdy jest tylko próżnia. Jak wiele dowcipów, również ten zawiera w sobie ziarno prawdy[60]. W szczególności prawdą jest, że oddziaływania grawitacyjne zaledwie trzech ciał przebiegające zgodnie z prawidłami newtonowskiego prawa odwrotnych kwadratów były przez całe stulecia niemożliwym do rozwiązania problemem matematycznym. Wciąż nim są, jeśli zależy nam na uzyskaniu eleganckiego wzoru na orbity ciał. Prawdę mówiąc, wiemy obecnie, że dynamika ruchu trzech ciał ma charakter chaotyczny – jest tak nieregularna, że zawiera w sobie elementy przypadkowości.

Wszystko to stoi w jaskrawym kontraście z niezwykłym wprost sukcesem newtonowskiego prawa powszechnego ciążenia, które pozwoliło między innymi odpowiedzieć na pytanie, jaki kształt mają orbity planet krążących wokół Słońca. Otrzymana odpowiedź zgadza się z wnioskiem, do jakiego doszedł wcześniej Kepler na podstawie obserwacji astronomicznych Marsa: orbity są elipsami. W tym wypadku mamy do czynienia tylko z dwoma ciałami – są nimi Słońce i planeta. Naturalnym następnym krokiem byłoby wykorzystanie

praw grawitacji Newtona do znalezienia równania opisującego orbity trzech ciał – i rozwiązanie go. Okazuje się jednak, że nie istnieje żaden elegancki opis geometrii orbit trzech ciał, ba, nie można nawet wyprowadzić wzoru, który opisywałby je na gruncie geometrii analitycznej. Do końca XIX stulecia uczeni nie potrafili zbyt wiele powiedzieć na temat ruchu trzech ciał niebieskich, nawet gdy jedno z nich jest tak małe, że jego masę można pominąć.

Od tego czasu udało nam się wiele dowiedzieć na temat dynamiki trzech (i więcej) ciał. Ważnym etapem tego postępu było uświadomienie sobie, jak bardzo trudny jest to problem i dlaczego tak jest. Mogłoby się wydawać, że oznacza to cofnięcie się w rozwoju, ale czasami najlepszym sposobem na dokonanie postępów jest zrobienie kroku wstecz i wypróbowanie innego podejścia. W przypadku problemu trzech ciał taki plan działania doprowadził do sukcesu, podczas gdy mozolne parcie do przodu zakończyłoby się zapewne zabrnięciem w ślepy zaułek.

Już ludzie pierwotni musieli zwrócić uwagę na to, że Księżyc przesuwa się stopniowo po nocnym niebie, przemieszczając się względem świecących w tle gwiazd. Gwiazdy również zdają się przemieszczać, ale robią to zespołowo, niczym maleńkie plamki światła na olbrzymiej obracającej się czaszy. Księżyc jest bez wątpienia wyjątkowy również pod innym względem – jest dużym świecącym kołem, którego kształt zmienia się od nowiu do pełni i z powrotem. Nie jest plamką światła, taką jak gwiazda.

Niektóre z tym plamek świetlnych zachowują się jednak inaczej. Błądzą po niebie. Nie zmieniają swojego położenia względem gwiazd tak szybko jak Księżyc, ale i tak nie trzeba wcale obserwować nieba przez wiele nocy, żeby się przekonać, że niektóre świecące punkty się poruszają. Pięć z tych błądzących gwiazd można dostrzec gołym okiem. Grecy nazwali je *planetes* – wędrowcy. Chodzi oczywiście o planety. Wspominane pięć planet, które znamy już od starożytności, to Merkury, Wenus, Mars, Jowisz i Saturn – wszystkie nazwane imionami rzymskich bogów. Za pomocą teleskopów udało się odkryć jeszcze dwie: Urana i Neptuna. No

i oczywiście musimy dodać do tego naszą Ziemię. Za sprawą kontrowersyjnej decyzji podjętej przez Międzynarodową Unię Astronomiczną w 2006 roku Pluton nie jest już uważany za planetę.

Starożytni filozofowie, astronomowie i matematycy badali uważnie niebo i uświadomili sobie, że planety nie błądzą po nim w sposób przypadkowy. Odkryli, że poruszają się one po krętych, ale przewidywalnych drogach i powracają w to samo miejsce na nocnym niebie w dosyć regularnych odstępach czasu. Obecnie wiemy, że kreślone przez nie wzory wynikają z ich ruchu okresowego po orbicie zamkniętej i z niewielkiego wkładu, jaki wnosi do ich powstawania ruch orbitalny samej Ziemi. Ustaliliśmy również, że okresowość tego ruchu nie jest dokładna – ale odstępstwa te są nieduże. Merkury potrzebuje niecałych 88 dni, żeby okrążyć Słońce, natomiast Jowiszowi zabiera to prawie 12 lat. Im dalej od Słońca znajduje się dana planeta, tym więcej czasu potrzebuje na pokonanie pełnej orbity.

Pierwszym modelem ruchu planet opisującym go dokładnie pod względem ilościowym był układ ptolemejski, nazwany tak od nazwiska Klaudiusza Ptolemeusza. Ptolemeusz opisał swój model w dziele *Almagest* (*Wielka rozprawa astronomiczna*) około 150 roku n.e. Jest to model geocentryczny – z Ziemią w środku – w którym wszystkie ciała niebieskie krążą wokół naszej planety. Poruszają się one tak, jak gdyby były przyczepione do całego szeregu gigantycznych sfer obracających się w stałym tempie wokół osi, która sama może być z kolei przytwierdzona do innej sfery. Takie połączenie wielu obracających się sfer było konieczne, żeby dało się przedstawić złożony ruch planet za pomocą uważanego za kosmiczny ideał jednostajnego ruchu po okręgu – czyli ruchu wokół równika sfery. Jeśli wprowadzi się wystarczająco dużo sfer i odpowiednio dobierze ich prędkości oraz układ osi, to model taki może bardzo dobrze oddawać rzeczywistość.

Mikołaj Kopernik zmodyfikował układ Ptolemeusza na kilka sposobów. Najbardziej radykalnym posunięciem było przyjęcie, że wszystkie ciała oprócz Księżyca obracają się wokół Słońca, a nie Ziemi – to doprowadziło do znacznego uproszczenia opisu. Tak powstał model heliocentryczny. Taka propozycja nie przypadła do

gustu Kościołowi katolickiemu, ale ostatecznie pogląd naukowy zwyciężył i wykształceni ludzie przyjęli do wiadomości, że to Ziemia obraca się wokół Słońca. W 1596 roku Kepler bronił układu kopernikańskiego w swoim dziele *Tajemnica kosmosu*, którego najważniejszym punktem było opisanie matematycznego związku między odległością planety od Słońca a jej prędkością orbitalną. Ustalił, że w miarę jak oddalamy się od Słońca, wzrost okresu orbitalnego kolejnych planet jest dwukrotnie większy od przyrostu odległości. Później doszedł do wniosku, że ten związek jest zbyt mało dokładny, by mógł być prawdziwy, ale właśnie ta zależność była pierwszym krokiem na drodze do opracowania dokładniejszego opisu. Kepler wytłumaczył również rozmieszczenie planet, wykorzystując do tego pięć wielościanów foremnych umieszczonych elegancko jeden w drugim i rozdzielonych obejmującymi je sferami. Fakt istnienia pięciu wielościanów foremnych wyjaśniał, dlaczego jest tylko pięć planet, ale obecnie przyjmuje się, że jest ich osiem, a zatem takie uzasadnienie przestało być pomocne. Istnieje 120 różnych sposobów ułożenia po kolei pięciu wielościanów foremnych i tylko jeden z nich daje proporcje zbliżone do proporcji zaobserwowanych na niebie orbit planet. Jest to więc tylko przypadkowa zbieżność i tego typu wyjaśnienie oznaczałoby wtłaczanie natury na siłę w układ pozbawiony jakiegokolwiek sensu.

W 1600 roku astronom Tycho Brahe zatrudnił Keplera, by pomógł mu w analizie obserwacji, ale plany ich współpracy pokrzyżowały problemy polityczne. Po śmierci Brahego Kepler został mianowany cesarskim matematykiem na dworze Rudolfa II. W wolnych chwilach analizował przeprowadzone przez Brahego obserwacje Marsa. Jednym z efektów tych prac było dzieło *Astronomia Nova* (*Nowa astronomia*) z 1609 roku, w którym Kepler przedstawił kolejne dwa prawa ruchu planet. Jego pierwsze prawo głosi, że planety poruszają się po elipsach – stwierdził, że tak właśnie jest w przypadku Marsa i wydawało się prawdopodobne, że również pozostałe orbity tak wyglądają. Początkowo zakładał, że do zebranych danych najlepiej będzie pasował kształt jajka, ale to się nie udało, sięgnął więc po elipsę. Również tę krzywą musiał odrzucić i ostatecznie znalazł inny

opis matematyczny kształtu orbity. Potem jednak zorientował się, że w istocie inną drogą wrócił do elipsy:

> Odłożyłem [uzyskane równanie] na bok i powróciłem do elips, sądząc, że jest to zupełnie inna hipoteza, podczas gdy, jak dowiodę w następnym rozdziale, są one identyczne [...]. Ach, jaki był ze mnie głuptas![61]

Drugie prawo Keplera stwierdza, że planety zakreślają równe obszary w równych odstępach czasu. W 1619 roku w książce *Harmonices Mundi* (*Harmonie świata*) uczony zamieścił swoje ostatnie, trzecie prawo, które jest dużo bardziej dokładnym opisem związku między odległościami i okresami orbitalnymi. Głosi ono, że sześcian odległości (połowy długości wielkiej osi elipsy) jest proporcjonalny do kwadratu okresu.

W tym momencie wszystko już było przygotowane na pojawienie się Isaaca Newtona. W swoim dziele *Matematyczne zasady filozofii przyrody* z 1687 roku Newton udowodnił, że trzy prawa Keplera są równoważne jednemu prawu grawitacji: dwa ciała przyciągają się z siłą proporcjonalną do ich mas i odwrotnie proporcjonalną do kwadratu dzielącej je odległości. Prawo Newtona miało olbrzymią przewagę: można je było stosować do dowolnego układu ciał, bez względu na ich liczbę. Ceną, jaką przyszło za to zapłacić, był sposób, w jaki opisuje ono orbity – nie za pomocą kształtów geometrycznych, ale w postaci rozwiązań równań różniczkowych zawierających przyspieszenia planet. Nie jest wcale oczywiste, w jaki sposób można na podstawie takich równań określić kształt orbit planet lub ich położenie w wybranej chwili. Mówiąc szczerze, nie jest nawet zbyt jasne, co trzeba zrobić, żeby wyznaczyć ich przyspieszenie. Niemniej równanie to w sposób *niejawny* zawiera tę informację. Problem polegał na wyrażeniu jej w sposób jawny. Keplerowi udało się tego dokonać dla dwóch ciał i otrzymał odpowiedź, że poruszają się one po orbitach eliptycznych z takimi prędkościami, iż w równych odstępach czasu zakreślają równe obszary.

A jeśli mamy trzy ciała?

Dobre pytanie. Zgodnie z prawem Newtona wszystkie ciała Układu Słonecznego wpływają na siebie grawitacyjnie. Dokładnie rzecz biorąc, należałoby raczej powiedzieć, że wszystkie ciała we Wszechświecie wpływają na siebie grawitacyjnie, ale nikt przy zdrowych zmysłach nie będzie próbował zapisać równań różniczkowych dla każdego ciała we Wszechświecie. Jak zwykle, wyjściem z tej sytuacji może być uproszczenie problemu – byle nie za daleko idące. Gwiazdy znajdują się tak daleko, że ich wpływ grawitacyjny na Układ Słoneczny jest zaniedbywalnie mały, z wyjątkiem przypadku, gdy chcemy opisać ruch Słońca w wirującej Galaktyce. Na ruch Księżyca wpływ mają głównie dwa inne ciała: Ziemia i Słońce, z wyjątkiem kilku niewielkich efektów wynikających z oddziaływania innych planet. Na początku XVIII wieku problem ten przestał być tylko pytaniem astronomicznym – stał się sprawą o znaczeniu praktycznym, ponieważ uświadomiono sobie, że ruch Księżyca można wykorzystać do nawigacji. (Nie było wówczas urządzeń do nawigacji satelitarnej, ani nawet chronometrów pozwalających zmierzyć długość geograficzną). Taka metoda wymagała jednak dokładniejszych przewidywań od tych, jakie dawały istniejące teorie. Było oczywiste, że należy wyjść od zapisania wniosków wynikających z prawa Newtona dla trzech ciał, które w tym wypadku można traktować jako masy punktowe, ponieważ planety są niezwykle małe w porównaniu z dzielącymi je odległościami. Potem należało rozwiązać otrzymane równania różniczkowe. Okazało się jednak, że sztuczki, za pomocą których Kepler uzyskał elipsy dla dwóch ciał, przestają działać, gdy dorzuci się jeszcze jedno, dodatkowe ciało. Kilka wstępnych kroków dało się przeprowadzić, ale potem uczeni nieodmiennie natrafiali na przeszkodę. W 1747 roku dwaj zawzięci rywale, Jean d'Alembert i Alexis Clairaut, wzięli udział w konkursie Paryskiej Akademii Nauk na rozwiązanie problemu trzech ciał. Obaj postanowili stawić czoło temu wyzwaniu, stosując przybliżenia liczbowe. O problemie trzech

ciał zrobiło się głośno i wkrótce uznano go za jedną z największych zagadek matematyki.

Udało się rozwiązać niektóre szczególne przypadki. W 1767 roku Euler odkrył rozwiązania, w których wszystkie trzy ciała leżą na obracającej się linii prostej. W 1772 roku Lagrange znalazł podobne rozwiązania dla ciał tworzących obracający się trójkąt równoboczny, który rozszerza się i kurczy. Oba rozwiązania są okresowe – ciała powtarzają w nich bez końca tę samą sekwencję ruchów. Jednak nawet daleko idące uproszczenia nie pozwoliły na uzyskanie jakiegokolwiek bardziej ogólnego rozwiązania. Można sobie zakładać, że jedno z ciał ma pomijalną masę czy że dwa pozostałe poruszają się po okręgach wokół wspólnego środka masy (ta wersja znana jest jako „ograniczony" problem trzech ciał), a równań i tak nie uda się rozwiązać analitycznie.

W latach 1860 i 1867 astronom i matematyk Charles-Eugène Delaunay zajął się konkretnym przypadkiem układu Słońce–Ziemia–Księżyc z wykorzystaniem rachunku zaburzeń, który pozwala przedstawić wpływ grawitacji Słońca na Księżyc jako niewielką poprawkę wnoszoną do wpływu, jaki wywiera nań Ziemia. Delaunay otrzymał w ten sposób przybliżony wzór w postaci szeregu, czyli sumy wielu kolejnych wyrazów. Wyniki swoich prac opublikował w latach 1860 i 1867 w postaci dwóch opasłych tomów liczących po 900 stron i zawierających głównie same wzory. Pod koniec lat siedemdziesiątych XX wieku sprawdzono jego obliczenia na komputerze i udało się w nich znaleźć tylko dwa niewielkie i nieistotne błędy.

Przeprowadzenie takich rachunków było nie lada wyczynem, ale szereg Delaunaya zbliża się do wartości granicznej zbyt wolno, by metoda ta mogła znaleźć praktyczne zastosowanie. Prace te zainspirowały jednak innych uczonych do poszukiwania takich rozwiązań, które będą zbiegały szybciej. Pozwoliły również odkryć dużą przeszkodę rachunkową w stosowaniu tego rodzaju metod, którą okazał się tak zwany problem małych mianowników. Chodzi o to, że pewne wyrazy w szeregu są ułamkami, których mianowniki stają się bardzo małe, jeśli ciała znajdują się w stanie rezonansu lub blisko niego. Przez rezonans rozumiemy tu ruch okresowy, w którym

okresy ruchu poszczególnych ciał są swoimi wymiernymi wielokrotnościami – na przykład trzy najbardziej wewnętrzne księżyce Jowisza, Io, Europa i Ganimedes, krążą wokół swojej macierzystej planety z okresami wynoszącymi 1,77, 3,55 i 7,15 dnia, a więc są one w niemal dokładnej proporcji 1:2:4. Szczególnie kłopotliwe są rezonanse precesji, czyli tempa obracania się osi dwóch orbit niemal eliptycznych, ponieważ nawet niewielki błąd w oszacowaniu ułamka staje się bardzo duży, gdy jego mianownik jest mały.

Skoro problem trzech ciał okazał się tak trudny, to z pewnością problem *n* ciał – opisu dowolnej liczby mas punktowych poruszających się zgodnie z newtonowskim prawem grawitacji – musi być jeszcze trudniejszy. Jednak natura przygotowała dla nas ważny przykład – jest nim cały Układ Słoneczny. Znajduje się w nim osiem planet, kilka planet karłowatych, takich jak Pluton, i tysiące asteroid, z których wiele osiąga całkiem spore rozmiary. Do tego moglibyśmy jeszcze dorzucić satelity, które czasami – na przykład Tytan – są większe od Merkurego. Zatem Układ Słoneczny może być problemem 10, 20 lub 1000 ciał, w zależności od tego, na jaki poziom szczegółowości opisu się zdecydujemy.

W wypadku przewidywań krótkoterminowych w zupełności wystarczają wartości przybliżone wyznaczone metodami numerycznymi, a w astronomii tysiąc lat to przecież bardzo krótko. Jeżeli jednak chcemy się dowiedzieć, jak będzie ewoluował Układ Słoneczny w ciągu setek milionów lat, to mamy do czynienia z czymś zupełnie innym. Taka długoterminowa prognoza potrzebna jest, by poznać odpowiedź na bardzo ważne pytanie: Czy Układ Słoneczny jest stabilny? Wydaje się, że planety poruszają się po względnie stabilnych, prawie eliptycznych orbitach. Orbity te zmieniają się nieznacznie za sprawą oddziaływań innych planet i w efekcie okres orbitalny może się zmienić o ułamek sekundy lub może się okazać, że wielkość elipsy nie jest do końca stała. Czy możemy mieć pewność, że w przyszłości ciała Układu Słonecznego też będą się tylko tak niewinnie popychać? Czy tak samo wyglądało to również

w przeszłości, szczególnie na wczesnym etapie formowania się naszego układu planetarnego? Czy Układ Słoneczny pozostanie stabilny, czy też może dojdzie w nim do zderzenia dwóch planet? Czy jakaś planeta może zostać wyrzucona w bezkresną otchłań Wszechświata?

W 1889 roku wypadały sześćdziesiąte urodziny Oskara II, króla Norwegii i Szwecji. Norweski matematyk Gösta Mittag-Leffler namówił monarchę, żeby dla uczczenia tego jubileuszu ufundował nagrodę za rozwiązanie problemu *n* ciał. Nie chodziło o przedstawienie dokładnego wzoru – wiedziano już wtedy, że byłyby to zbyt wygórowane oczekiwania – ale jakiegoś zbieżnego szeregu. Wyzwanie to zainteresowało Poincarégo, który postanowił zacząć od bardzo prostej wersji problemu. Zajął się więc ograniczonym problemem trzech ciał, z których jedno ma zaniedbywalnie małą masę, niczym maleńki pyłek. Jeśli do takiego pyłku zastosujemy bezpośrednio prawo Newtona, to stwierdzimy, że działająca na niego siła jest proporcjonalna do iloczynu mas podzielonego przez kwadrat odległości, a ponieważ jedna z mas jest zerowa, to również cały iloczyn jest równy zeru. Takie rozwiązanie nie jest zbyt pomocne, ponieważ oznacza, że pyłek po prostu odleci w dal, niezwiązany z pozostałymi dwoma ciałami żadną siłą. Zamiast tego należy więc przygotować model, w którym pyłek odczuwa wpływ obydwu dużych ciał, natomiast one zupełnie nie odczuwają wywieranej przez niego siły. W takim razie obydwa masywne ciała poruszają się po orbitach kołowych ze stałą prędkością. Cała złożoność ruchu zostaje przekazana cząsteczce pyłu.

Poincaré nie rozwiązał problemu, który postawił król Oskar. Ten cel okazał się zbyt ambitny. Jednak jego metody były tak nowatorskie i udało mu się dokonać tak dużego postępu, że mimo to przyznano mu obiecaną nagrodę. Jego zwycięską pracę opublikowano w 1890 roku i wynikało z niej, że nawet dla ograniczonego problemu trzech ciał rozwiązanie w postaci wymaganej w konkursie może być niemożliwe do uzyskania. Poincaré podzielił analizę na kilka różnych przypadków, w zależności od ogólnych cech ruchu. W większości z nich uzyskanie rozwiązania w postaci szeregu powinno być

możliwe. Jednak w jednym przypadku orbita cząsteczki pyłu stawała się nadzwyczaj nieregularna.

Poincaré domyślił się istnienia tej nieuniknionej nieregularności na podstawie innych idei, które w tym czasie badał. Idee te pozwoliły mu na opisanie rozwiązań równań różniczkowych, bez konieczności ich faktycznego rozwiązywania. Ta „jakościowa teoria równań różniczkowych" okazała się zarodkiem, z którego rozwinęła się współczesna dynamika nieliniowa. Podstawowy pomysł polegał na zbadaniu geometrii rozwiązań, a właściwie – ich topologii. (Poincaré bardzo interesował się również zagadnieniami topologicznymi – powiemy o tym w rozdziale 10). W takim ujęciu położenia i prędkości ciał stają się współrzędnymi w wielowymiarowej przestrzeni. W miarę upływu czasu każdy stan początkowy przemierza w tej przestrzeni jakąś trajektorię. To właśnie topologia tej trajektorii, lub całego układu wszystkich możliwych trajektorii, pozwala nam się domyślić wielu ważnych rzeczy na temat rozwiązań.

Weźmy na przykład rozwiązanie okresowe, które tworzy trajektorię w kształcie zamkniętej pętli. W miarę upływu czasu stan układu przemieszcza się wzdłuż tej pętli, powtarzając bez końca takie samo zachowanie. Badany układ jest w takim wypadku okresowy. Poincaré zaproponował, że dobrym sposobem na wykrywanie takich pętli mogłoby być użycie wielowymiarowej powierzchni, umieszczonej w taki sposób, by przecinała pętlę. Obecnie powierzchnie takie nazywamy przekrojami Poincarégo. Rozwiązania wybiegające z takiej powierzchni mogą ostatecznie do niej powrócić. Pętla powraca w tym samym punkcie, a rozwiązania przebiegające przez pobliskie punkty zawsze powracają do tego obszaru po mniej więcej jednym okresie. Zatem rozwiązanie okresowe można przedstawić jako stały punkt na „mapie pierwszego powrotu", z której wynika, co dzieje się z punktami na powierzchni, gdy wracają na nią po raz pierwszy, jeśli w ogóle wracają. Mogłoby się wydawać, że nie dokonaliśmy w ten sposób zbyt dużego postępu, ale takie rozwiązanie pozwala zmniejszyć wymiar przestrzeni, czyli liczbę analizowanych zmiennych. To niemal zawsze się opłaca.

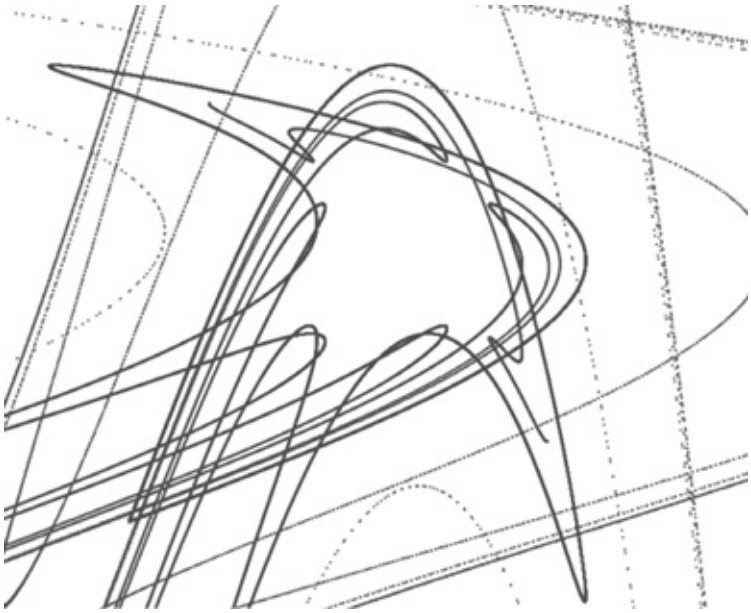
Prawdziwą wielkość pomysłu Poincarégo docenimy jednak w pełni dopiero wtedy, gdy przejdziemy do bardziej skomplikowanego

rodzaju rozwiązania, które jest połączeniem kilku ruchów okresowych. Jako prosty przykład rozważmy Ziemię okrążającą Słońce mniej więcej w ciągu 365 dni i Księżyc, który wykonuje jedno pełne okrążenie Ziemi w przybliżeniu co 28 dni. W takim układzie ruch Księżyca jest połączeniem dwóch różnych ruchów okresowych. Oczywiście cały sens rozważania układów trzech ciał polega na tym, że ów opis nie jest w pełni dokładny, ale takie „półokresowe" rozwiązania pojawiają się bardzo często w układach wielu ciał. Przekrój Poincarégo pozwala wykryć również takie półokresowe rozwiązania. Poznajemy je po tym, że gdy wracają na powierzchnię, nie trafiają w to samo miejsce, ale punkt, w którym za każdym razem przecinają powierzchnię, przesuwa się powoli po krzywej zamkniętej leżącej *na tej powierzchni*.

Poincaré uświadomił sobie, że gdyby w taki sposób zachowywało się każde rozwiązanie, to można by utworzyć odpowiednie szeregi opisujące je ilościowo. Gdy jednak przeanalizował topologię mapy pierwszego powrotu, zauważył, że sytuacja może być jeszcze bardziej skomplikowana. Otóż dwie określone krzywe, związane ze sobą za sprawą dynamiki, mogą się przecinać. Fakt ten, sam w sobie, nie musi jeszcze oznaczać nic złego, ale gdy się przedłuży te krzywe aż do miejsca, w którym ponownie docierają do powierzchni, to znowu muszą się przeciąć – teraz jednak w innym miejscu. Jeśli przedłużymy je o jeszcze jeden obrót, to przetną się kolejny raz. Cała sprawa gmatwa się jeszcze bardziej: te nowe krzywe, powstałe przez przedłużanie krzywych pierwotnych, nie są tak naprawdę wcale nowe. Są częścią krzywych pierwotnych. Zrozumienie topologii tej plątaniny wymagało spokojnego przemyślenia wszystkiego, ponieważ nikt wcześniej nie zajmował się jeszcze tego typu zagadnieniami. Powstał w ten sposób bardzo skomplikowany obraz przypominający jakąś zwariowaną sieć, w której linie wiją się wielokrotnie tu i tam, potem nagle się przecinają i znowu wyginają w szalone zawijasy i tak dalej, z dowolnym stopniem złożoności. Poincaré w zasadzie przyznał, że zupełnie zbiło go to z tropu:

> Gdy próbuje się opisać figurę tworzoną przez te krzywe i nieskończoną ilość ich punktów przecięcia, z których każdy odpowiada podwójnie asymptotycznemu rozwiązaniu, to punkty te tworzą jakiś rodzaj siatki, pajęczyny, a może nieskończenie gęstej sieci […] i nie można się oprzeć zdumieniu złożonością tej figury, której nawet nie będę próbował rysować.

Obecnie plątaninę tę nazywamy splotem homoklinicznym („połączony z samym sobą”; zob. ryc. 31). Dzięki nowym ideom topologicznym, które Stephen Smale wprowadził do matematyki w latach sześćdziesiątych XX wieku, rozpoznajemy teraz w tej strukturze starego znajomego. Najważniejszym wnioskiem płynącym z jej badań jest to, że taka dynamika jest *chaotyczna*. Chociaż w wyjściowych równaniach nie ma żadnego jawnego elementu przypadkowości, ich rozwiązania są bardzo skomplikowane i nieregularne, mają więc pewne cechy wspólne z procesami naprawdę losowymi. Istnieją na przykład orbity – a tak naprawdę, to tak wygląda większość z nich – których ruch dokładnie naśladuje wielokrotny, losowy rzut monetą. Odkrycie, że układ deterministyczny – czyli taki, którego cała przyszłość jednoznacznie wynika z jego stanu obecnego – może jednak przejawiać cechy przypadkowości, jest niezwykłe i odmieniło wiele dziedzin nauki. Nie zakładamy już automatycznie, że proste reguły prowadzą do prostego zachowania. Można zatem powiedzieć, że ten obszar badań, nazywany potocznie teorią chaosu, powstał dzięki pracom Poincarégo i jego udziałowi w konkursie ogłoszonym przez króla Oskara.



Ryc. 31. Fragment splotu homoklinicznego. Pełny rysunek byłby nieskończenie skomplikowany.

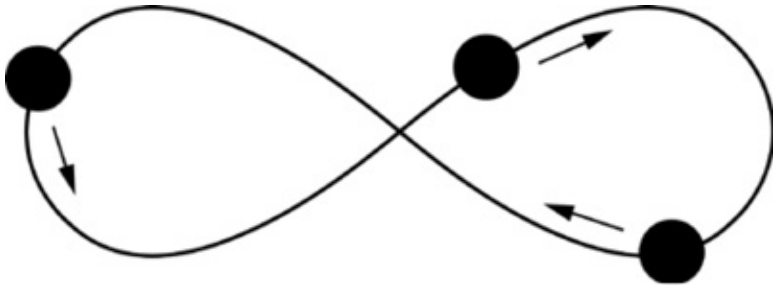
No, prawie. Przez wiele lat historycy i matematycy tak właśnie relacjonowali tę historię. Jednak około roku 1990 June Barrow-Green natknęła się w przepastnych archiwach Instytutu Mittaga-Lefflera w Sztokholmie na wydany drukiem egzemplarz rozprawy Poincarégo, przejrzała go pobieżnie i zdała sobie sprawę, że ta wersja różni się od pracy, którą można znaleźć we wszystkich bibliotekach matematycznych na całym świecie. Było to oficjalne wydanie pracy, którą Poincaré zgłosił do konkursu, i okazało się, że jest w niej błąd. W rozprawie przesłanej na konkurs bowiem nie uwzględnił rozwiązań chaotycznych. Zdążył zauważyć swoją pomyłkę jeszcze przed wydaniem dzieła drukiem, ustalił, co z tych rozwiązań wynika – oczywiście: chaos – i zapłacił (więcej, niż wynosiła nagroda) za zniszczenie oryginalnej wersji i wydrukowanie poprawionej. Z jakiegoś powodu w archiwach Instytutu Mittaga-Lefflera przechowano egzemplarz pierwotnej, wadliwej wersji, ale nikt o tym nie pamiętał i dopiero June Barrow-Green wydobyła tę historię na światło dzienne, opisując swoje znalezisko w książce wydanej w 1994 roku.

Wydaje się, że Poincaré uważał, iż takich chaotycznych rozwiązań nie można przedstawić w postaci rozwinięcia w szereg, ale okazuje się, że również w tej kwestii się pomylił. Nietrudno było dojść do takiego wniosku, bo przecież szeregi wydają się zbyt regularne, by mogły opisywać chaos. Jedynie topologia daje takie możliwości. Chaos jest skomplikowanym zachowaniem wynikającym z prostych reguł, zatem taki wniosek nie jest niepodważalny, ale struktura problemu trzech ciał z pewnością uniemożliwia uzyskanie prostych rozwiązań podobnych do tych, jakie Newton otrzymał dla dwóch ciał. Problem dwóch ciał jest „całkowalny", co oznacza, że równania zawierają wystarczająco dużo zachowywanych wielkości, takich jak energia, pęd i moment pędu, by można było wyznaczyć orbity. Wielkości „zachowywane" to takie, które nie zmieniają się, gdy ciała podążają po swoich orbitach. O problemie trzech ciał wiadomo natomiast, że nie jest całkowalny.

Mimo to istnieją rozwiązania mające postać szeregów, ale nie zachowują one poprawności we wszystkich sytuacjach. Załamują się dla stanów początkowych o zerowym momencie pędu – który jest miarą całkowitego obrotu ciała – ale takie przypadki zdarzają się niesłychanie rzadko, ponieważ zero jest tylko jedną liczbą wśród nieskończonej ilości wszystkich liczb rzeczywistych. Poza tym nie są to szeregi samej zmiennej czasowej, ale jej pierwiastka sześciennego. Fiński matematyk Karl Frithiof Sundman odkrył ten fakt w 1912 roku. Z pracy Qiudonga Wanga opublikowanej w 1991 roku wynika natomiast, że podobna sytuacja występuje dla problemu *n* ciał, także z pewnymi rzadkimi wyjątkami, ale w przypadku czterech lub więcej ciał nie potrafimy przedstawić żadnej klasyfikacji dokładnych okoliczności, w których szeregi przestają być zbieżne. Wiemy, że taka klasyfikacja musi być bardzo skomplikowana, ponieważ istnieją rozwiązania, w których wszystkie ciała rozbiegają się na nieskończenie dużą odległość lub po pewnym skończonym czasie wpadają w ruch oscylacyjny o nieskończonej prędkości (więcej na temat powiemy w rozdziale 12). Z punktu widzenia fizyki można powiedzieć, że takie rozwiązania są wynikiem przyjęcia założenia, iż ciała są pojedynczymi (masywnymi) punktami. Na gruncie

matematyki podpowiadają nam one jednak, w jakich obszarach układ może się zachowywać w sposób szalony.

Uczonym udało się mimo to dokonać znacznych postępów w analizie problemu *n* ciał dla przypadku, gdy wszystkie ciała mają jednakową masę. Takie założenie rzadko kiedy jest realistyczne w odniesieniu do ciał niebieskich, ale wydaje się zupełnie sensowne w badaniach niektórych niekwantowych modeli cząstek elementarnych. Tego typu analizy są jednak najbardziej interesujące z punktu widzenia matematyki. W 1993 roku Christopher Moore znalazł rozwiązanie problemu trzech ciał, w którym ciała te gonią się wzajemnie po jednej orbicie. Jeszcze bardziej zaskakujący jest wygląd tej orbity – ma ona mianowicie kształt cyfry 8 (zob. ryc. 32). Chociaż taka krzywa przecina samą siebie, poruszające się po niej ciała nigdy się nie zderzają.



Ryc. 32. Choreografia w kształcie cyfry 8

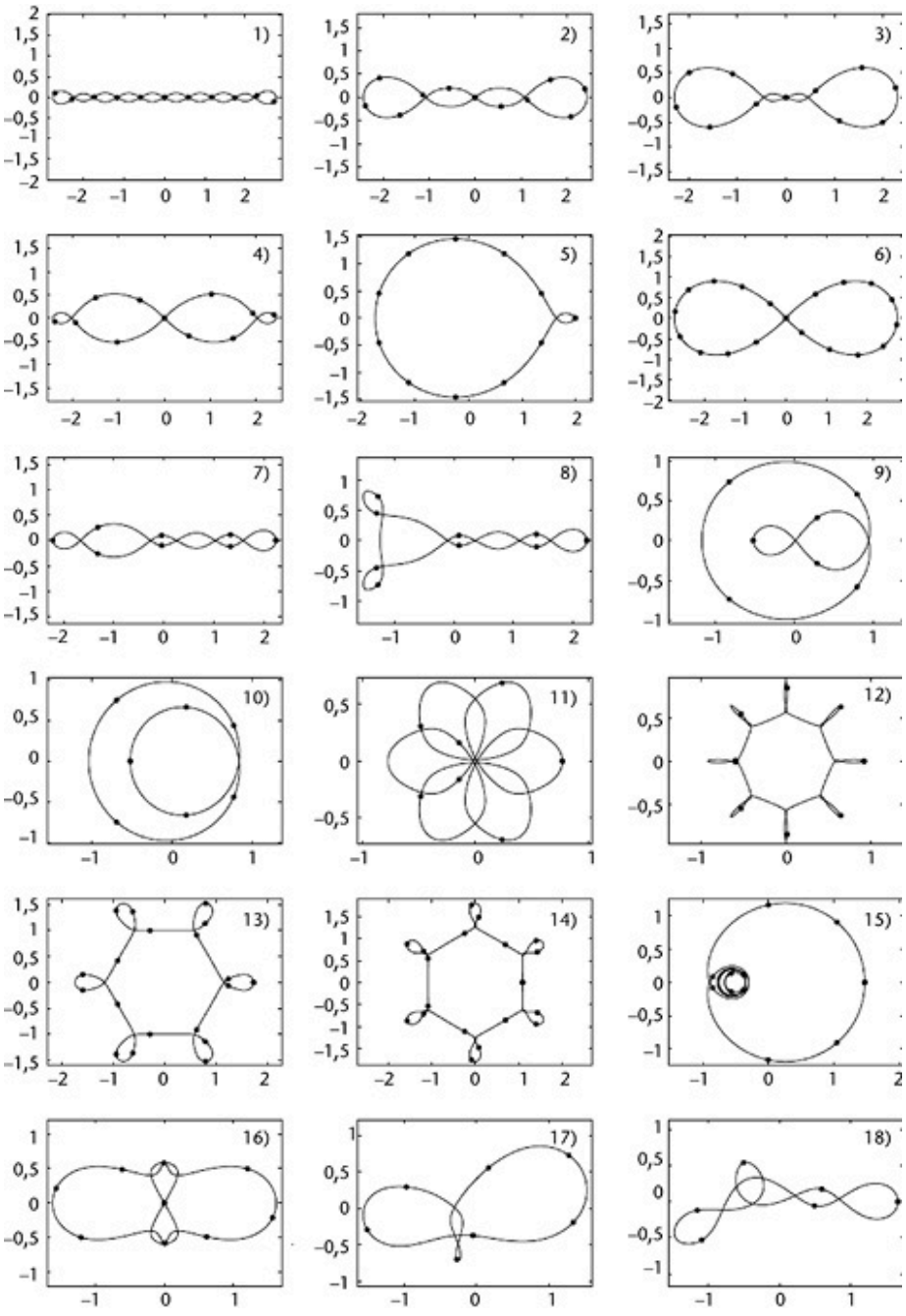
Moore przeprowadził swoje obliczenia numerycznie, na komputerze. To samo rozwiązanie odkryli niezależnie w 2001 roku Alain Chenciner i Richard Montgomery, którzy wykorzystali uświęconą tradycją zasadę mechaniki klasycznej, znaną jako „zasada najmniejszego działania", i dość skomplikowaną topologię do przedstawienia ścisłego dowodu, że takie rozwiązanie musi istnieć. Orbity te są okresowe, co oznacza, że wszystkie ciała powtarzają bez końca te same ruchy i po upływie stałego odstępu czasu będą miały takie same położenia i prędkości jak na początku. Przy ustalonej,

jednakowej masie wszystkich ciał istnieje przynajmniej jedno takie rozwiązanie dla każdego okresu.

W 2000 roku Carles Simó przeprowadził analizę numeryczną, z której wynika, że taka orbita w kształcie ósemki jest stabilna i może w niej dochodzić najwyżej do bardzo powolnego, długoterminowego dryfu, znanego jako dyfuzja Arnolda, związanego ze szczegółami geometrii mapy powrotu Poincarégo. W wypadku tego rodzaju stabilności *prawie* wszystkie zaburzenia prowadzą do powstania orbity bardzo zbliżonej do rozważanej i w miarę jak stają się one coraz mniejsze, odsetek tak zachowujących się zaburzeń zbliża się do 100 procent. W wypadku tego niewielkiego odsetka zaburzeń, które nie zachowują się w tak stabilny sposób, orbita oddala się od swojego pierwotnego położenia niezwykle powoli. Wyniki uzyskane przez Simó były dla wszystkich zaskoczeniem, ponieważ stabilne orbity występują bardzo rzadko w problemie trzech ciał o jednakowych masach. Z obliczeń numerycznych wynika, że orbity te pozostają stabilne nawet wtedy, gdy masy ciał są nieznacznie różne. Nie jest więc wykluczone, że gdzieś we Wszechświecie trzy gwiazdy o niemal jednakowych masach gonią się po orbicie w kształcie cyfry 8. W 2000 roku Douglas Heggie oszacował, że liczba takich układów potrójnych gwiazd zawiera się w przedziale między jeden na galaktykę a jeden na cały Wszechświat.

Krzywa w kształcie cyfry 8 ma interesującą symetrię. Przyjmijmy, że nasze trzy ciała to: *A*, *B* i *C*. Pozwólmy im przemieszczać się po orbicie przez jedną trzecią okresu i zatrzymajmy symulację. Jeśli teraz spojrzymy na orbitę, przekonamy się, że znajdują się na niej trzy ciała o takich samych położeniach i prędkościach jak na początku, z tą jednak różnicą, że teraz ustawione są w kolejności *B*, *C* i *A*. Po dwóch trzecich okresu ponownie uzyskamy taki sam układ ciał jak na początku, ale teraz będą one ułożone na orbicie w kolejności *C*, *A*, *B*. Po upływie pełnego okresu orbitalnego otrzymamy taki sam układ i kolejność ciał jak na początku. Tego typu rozwiązanie nazywamy choreografią – przypomina ono planetarny taniec, w którym wszyscy co pewien czas zamieniają się miejscami. Z analiz numerycznych wynika, że istnieją również

choreografie większej liczby ciał – na rycinie 33 pokazano kilka ich przykładów. Szczególnie dużą liczbę różnych choreografii znalazł Simó[62].



Ryc. 33. Przykłady choreografii

Jednak nawet w tej sytuacji wiele pytań pozostaje bez odpowiedzi. Wciąż nie mamy ścisłego dowodu istnienia takich choreografii. Wydaje się, że w wypadku większej liczby ciał niż trzy powinny one być niestabilne. Zapewne tak jest, ale nikt tego jeszcze nie udowodnił. Wygląda na to, że orbita w kształcie ósemki dla trzech ciał o jednakowej masie jest pod tym względem wyjątkowa, ale również tego nie udało się jeszcze udowodnić, chociaż w 2003 roku Tomasz Kapela i Piotr Zgliczyński przedstawili wspomagany komputerowo dowód, że takie rozwiązanie jest lokalnie jednoznaczne – czyli że żadna leżąca w pobliżu orbita nie zadziała. Niewykluczone, że choreografie już niedługo okażą się kolejnym wielkim problemem.

No to jak jest z Układem Słonecznym – jest stabilny czy nie?

Może tak, a może nie.

Dzięki badaniom wykrytej przez Poincarégo możliwości pojawienia się chaosu udało nam się lepiej zrozumieć teoretyczne zagadnienia związane z osiągnięciem stabilności. Okazuje się, że są to problemy bardzo subtelne, złożone i – jak na ironię – w żaden istotny sposób niezwiązane z istnieniem rozwiązań, które można zapisać w postaci szeregów. Prace Jürgena Mosera i Władimira Arnolda pokazały, że różne uproszczone modele Układu Słonecznego są stabilne dla niemal wszystkich stanów początkowych, być może z wyjątkiem występowania efektu dyfuzji Arnolda, który uniemożliwia pojawienie się silniejszych rodzajów stabilności w niemal wszystkich tego rodzaju przypadkach. W 1961 roku Arnold udowodnił, iż wyidealizowany model układu planetarnego będzie w tym sensie stabilny, ale tylko przy założeniu, że planety mają bardzo małe masy w porównaniu z masą macierzystej gwiazdy, a ich orbity są bardzo zbliżone do orbit kołowych i leżą bardzo blisko jednej wspólnej płaszczyzny. Z przeprowadzonego dowodu wynika, że „bardzo blisko” i „bardzo zbliżone” oznacza tu „różniące się najwyżej o czynnik 10^{-43}” – nawet przy takich różnicach prawdopodobieństwo wystąpienia niestabilności będzie zerowe. W tego rodzaju rozważaniach perturbacyjnych wyniki zachowują zwykle swoją ważność nawet dla

znacznie większych rozbieżności, niż daje się to ściśle udowodnić, zatem można przyjąć, że układy planetarne dość bliskie tego ideału powinny być stabilne. Jednak dla naszego Układu Słonecznego rozbieżności te wynoszą około 10^{-3} dla mas i 10^{-2} dla kolistości i nachylenia orbit. To zdecydowanie więcej niż 10^{-43}, a więc pozostaje kwestią dyskusyjną, czy wyniki Arnolda mają tym wypadku zastosowanie. Niemniej możemy czerpać otuchę z faktu, że w ogóle udało się w tej kwestii *cokolwiek* stwierdzić z całkowitą pewnością.

Dzięki rozwojowi technik numerycznych pozwalających rozwiązywać równania metodami przybliżonymi z wykorzystaniem komputerów udało się również lepiej zrozumieć wiele praktycznych kwestii związanych z tego typu zagadnieniami. Takie procesy są bardzo wrażliwe, ponieważ zjawiska chaotyczne mają jedną ważną cechę: niewielkie błędy mogą bardzo szybko urosnąć do olbrzymich rozmiarów i zepsuć wynik. Teoretyczne badania chaosu i równań takich jak równania opisujące układ planetarny, w których nie uwzględnia się tarcia, pozwoliły na opracowanie metod numerycznych odpornych na wiele bardzo irytujących cech chaosu. Są to tak zwane symplektyczne metody całkowania. Jeśli za ich pomocą przeprowadzimy analizę orbity Plutona, to okaże się, że jest ona chaotyczna. Nie oznacza to jednak, że Pluton pędzi na oślep przez Układ Słoneczny, siejąc wkoło zniszczenie. Wynika z tego jedynie to, że za 200 milionów lat nadal będzie on znajdował się gdzieś blisko swej obecnej orbity, ale nie mamy najmniejszego pojęcia, gdzie dokładnie będzie ona przebiegała.

W 1982 roku poznaliśmy wyniki badań przeprowadzonych w ramach wymyślonego przez Archiego Roya projektu pod nazwą Longstop. Pracujący w tym projekcie uczeni zbadali na superkomputerze model planet zewnętrznych (obejmujący Jowisza i dalej położone planety) i nie wykryli żadnej wielkoskalowej niestabilności, chociaż zauważyli, że niektóre planety w bardzo dziwny sposób pozyskiwały energię kosztem innych. Później dwie inne grupy badaczy zainteresowały się rozwojem takich metod obliczeniowych i zastosowaniem ich do analizy wielu różnych kwestii związanych z Układem Słonecznym. Są to zespoły pracujące pod

kierunkiem Jacka Wisdoma i Jacques'a Laskara. W 1984 roku zespół Wisdoma przewidział, że obracający się wówczas dość regularnie księżyc Saturna – Hyperion – powinien chaotycznie koziołkować, co potwierdziły późniejsze obserwacje. W 1988 roku uczeni z tego zespołu we współpracy z Gerrym Sussmanem zbudowali własny komputer zaprojektowany specjalnie w celu rozwiązywania równań mechaniki nieba – było to prawdziwe cyfrowe planetarium. Uruchomiona na nim symulacja pozwoliła badaczom prześledzić zachowanie Układu Słonecznego w ciągu najbliższych 845 milionów lat i doprowadziła do odkrycia chaotycznej natury Plutona. Później badacze z grupy Wisdoma zbudowali kolejne tego typu komputery i dzięki nim udało się zbadać dynamikę Układu Słonecznego w okresie najbliższych kilku miliardów lat.

Zespół Laskara opublikował pierwsze wyniki badań długoterminowego zachowania Układu Słonecznego w 1989 roku. W swoich badaniach uczeni ci wykorzystali uśrednioną postać równań zaproponowanych jeszcze przez Lagrange'a, co sprawiło, że niektóre szczegóły uległy rozmyciu i nie były brane pod uwagę. Wyniki tych obliczeń sugerują, że położenie Ziemi na jej orbicie jest chaotyczne, podobnie jak Plutona. Oznacza to, że jeśli zmierzymy obecne położenie Ziemi i pomylimy się zaledwie o 15 metrów, to w żaden sposób nie będziemy mogli przewidzieć jej pozycji na orbicie za 100 milionów lat.

Jednym ze sposobów na zminimalizowanie wpływu chaosu jest przeprowadzenie wielu symulacji z nieznacznie różniącymi się danymi początkowymi. Można w ten sposób poznać cały zakres możliwych obrazów przyszłości wraz z prawdopodobieństwem zaistnienia każdego z nich. W 2009 roku Laskar i Mickaël Gastineau zastosowali taką technikę do badania Układu Słonecznego i przeanalizowali 2500 różnych scenariuszy. Różnice między nimi były niewiarygodnie małe – na przykład dla Merkurego w grę wchodziło przesunięcie o zaledwie 1 metr. Mniej więcej jeden procent tych scenariuszy przewiduje, że Merkury staje się w przyszłości niestabilny: zderza się z Wenus, spada na Słońce lub odlatuje w przestrzeń kosmiczną.

W 1999 roku Norman Murray i Matthew Holman zainteresowali się rozbieżnością między przewidywaniami takimi jak te wysunięte przez Arnolda, z których wynika stabilność układu planetarnego, a wynikami symulacji, wskazującymi na niestabilność. „Czy niepoprawne są wyniki symulacji numerycznych, czy może klasyczne obliczenia nie mają w tych wypadkach zastosowania?" – spytali. Korzystając z metod analitycznych – nie numerycznych – dowiedli, że w takich sytuacjach to jednak obliczenia klasyczne zawodzą. Właściwie odzwierciedlenie rzeczywistości wymaga uwzględnienia zbyt dużych zaburzeń. Głównym źródłem chaosu w Układzie Słonecznym jest istnienie układu niemal rezonansowego między Jowiszem, Saturnem i Uranem, a także drugiego, mniej ważnego, między Saturnem, Uranem i Neptunem. Murray i Holman sprawdzili ten wniosek, wykorzystując metody numeryczne, i ustalili, że horyzont predykcji – czyli wielkość będąca miarą czasu, jaki musi upłynąć, zanim małe błędy staną się na tyle duże, że zaczną wpływać na wynik – wynosi w tym wypadku około 10 milionów lat[63]. Przeprowadzone przez nich symulacje pokazują, że w wyniku chaotycznych zmian mimośrodu orbity Urana planeta ta od czasu do czasu zbliża się do Saturna i istnieje pewne ryzyko, iż w końcu zostanie wyrzucona poza obręb Układu Słonecznego. Na szczęście najbardziej prawdopodobne jest, że nastąpi to dopiero za około 10^{18} lat. Znacznie wcześniej, bo za jakieś 5 miliardów lat, Słońce wybuchnie i zmieni się w czerwonego olbrzyma, a to z pewnością wywrze wpływ na wszystkie planety, choćby dlatego, że Słońce utraci wówczas 30 procent swojej masy. Ziemia przesunie się bliżej zewnętrznej krawędzi Układu Słonecznego i być może uda jej się uniknąć wchłonięcia przez puchnące Słońce. Z tego, co obecnie wiemy, wynika jednak, że oddziaływania pływowe przyciągną ostatecznie Ziemię w kierunku Słońca. Znacznie wcześniej zagotują się i wyparują wszystkie oceany. Ponieważ teoria ewolucji wskazuje, że przeciętny czas istnienia gatunku nie przekracza 5 milionów lat, nie musimy się martwić żadnym z tych potencjalnie katastroficznych scenariuszy. Coś innego dopadnie nas wcześniej.

Te same metody można wykorzystać do zbadania przeszłości Układu Słonecznego. Wystarczy zastosować prostą matematyczną sztuczkę i użyć tych samych równań, cofając się po prostu w czasie. Do niedawna astronomowie zakładali, że planety zawsze znajdowały się blisko swoich obecnych orbit – że są na nich już od chwili, gdy powstały z obłoku gazu i pyłu otaczającego rodzące się Słońce. Ich orbity i skład chemiczny posłużyły nawet uczonym do wyznaczenia rozmiaru i składu owego pierwotnego obłoku pyłu. Obecnie jednak wydaje się, że planety nie pojawiły się od razu na swoich obecnych orbitach. Gdy obłok pyłu zapadał się pod wpływem własnej grawitacji, Jowisz – najbardziej masywna z planet – zaczął decydować o położeniu innych ciał niebieskich, a one z kolei wpływały wzajemnie na siebie. Taką hipotezę wysunęli jako pierwsi w 1984 roku Julio Fernandez i Wing-Huen Ip, ale przez pewien czas ich pracę uważano tylko za ciekawostkę. W 1993 roku Renu Malhotra zaczęła poważnie analizować, jak zmiany orbity Neptuna mogą wpływać na pozostałe gazowe olbrzymy, i wówczas inni uczeni zainteresowali się tym zagadnieniem. W ten sposób powstał obraz bardzo dynamicznego Układu Słonecznego.

W epoce gdy planety wciąż gromadziły materię, nastąpił taki okres, w którym Jowisz, Saturn, Uran i Neptun były już niemal gotowe, choć w ich pobliżu wciąż krążyła olbrzymia liczba skalistych i lodowych planetozymali – czyli małych ciał niebieskich o średnicy około 10 kilometrów. Od tej chwili dalsza ewolucja Układu Słonecznego była wynikiem migracji i zderzeń planetozymali. Wiele z nich wyleciało poza Układ Słoneczny, co zmniejszyło energię i moment pędu czterech gazowych olbrzymów. Ponieważ planety te miały różne masy i znajdowały się w różnych odległościach od Słońca, reagowały na te zdarzenia w różny sposób. Neptun okazał się jednym ze zwycięzców w rywalizacji o energię orbitalną i przesunął się na zewnątrz. Podobnie Uran i Saturn, choć w mniejszym stopniu. Wielkim przegranym, z punktu widzenia energii, był Jowisz, który przesunął się bliżej środka Układu Słonecznego. Był jednak tak masywny, że nie przemieścił się zbyt daleko.

Zmiany te wpłynęły również na pozostałe, mniejsze ciała Układu Słonecznego. Jego obecny, jak się wydaje, stabilny rozkład powstał w wyniku skomplikowanego tańca gazowych olbrzymów, w trakcie którego te gigantyczne planety ciskały w siebie mniejszymi ciałami, siejąc wkoło chaos.

No to jak jest z Układem Słonecznym – jest stabilny czy nie? Zapewne nie, ale my się o tym nie przekonamy, bo już dawno nas nie będzie.

60 A może nie. Władimir Kriwczenkow zwrócił uwagę na fakt, że energię stanu podstawowego i pierwszych stanów wzbudzonych kwantowego odpowiednika problemu trzech ciał można wyliczyć nawet ręcznie. Natomiast w mechanice klasycznej analogiczny problem jest bardziej skomplikowany za sprawą chaosu.

61 Arthur Koestler, *Lunatycy: historia zmiennych poglądów człowieka na wszechświat*, przeł. Tomasz Bieroń, Zysk i S-ka Wydawnictwo, Poznań 2002, s. 331 (przyp. tłum.).

62 Animacje choreografii i więcej informacji (w języku angielskim) można znaleźć na stronie internetowej: http://www.scholarpedia.org/article/N-body_choreographies.

63 Fachowo ten przedział czasu nazywamy czasem Lapunowa.

9. Jeszcze o liczbach pierwszych

Hipoteza Riemanna

W rozdziale 2 przyjrzeliśmy się własnościom poszczególnych liczb pierwszych i zauważyliśmy, że swoim zmiennym i nieprzewidywalnym zachowaniem przypominają istoty ludzkie. Ludzie mają wolną wolę, mogą dokonywać własnych wyborów, kierując się sobie tylko znanymi pobudkami. Liczby pierwsze muszą przestrzegać logiki, jaką narzuca im arytmetyka, ale również można odnieść wrażenie, że mają własną wolę. Ich zachowaniem rządzą dziwne zbiegi okoliczności i nierzadko wydaje się, że nie stoi za tym żadna sensowna struktura.

Niemniej w świecie liczb pierwszych nie króluje wcale anarchia. W 1835 roku Adolphe Quetelet zadziwił współczesnych odkryciem matematycznych prawidłowości w zdarzeniach społecznych wynikających ze świadomych decyzji ludzi lub działania losu – w danych dotyczących narodzin, związków małżeńskich, śmierci czy samobójstw. Były to dane statystyczne – nie dotyczyły konkretnych osób, ale uśrednionego zachowania dużych grup ludzi. Właśnie w taki sposób statystycy dostrzegają porządek w decyzjach ludzi kierujących się wolną wolą. Mniej więcej w tym samym okresie matematycy zaczęli zdawać sobie sprawę, że te same metody sprawdzają się w odniesieniu do liczb pierwszych. Chociaż każda z nich jest zatwardziałą indywidualistką, w grupie poddają się działaniu prawa. Można w nich dostrzec ukryte prawidłowości.

Wzorce statystyczne wyłaniają się wtedy, gdy rozważamy całe przedziały liczb pierwszych. Na przykład: ile jest liczb pierwszych mniejszych od określonej liczby? Na tak postawione pytanie bardzo

trudno jest udzielić dokładnej odpowiedzi, ale mamy doskonałe przybliżenia, które są tym bliższe prawdy, im większa jest owa wartość graniczna. W niektórych wypadkach potrafimy sprawić, że różnica między oszacowaniem a wartością dokładną staje się bardzo mała, ale zwykle nie jest to możliwe. Większość tego typu przybliżeń ma charakter asymptotyczny, co oznacza, że stosunek oszacowania do wartości dokładnej może stać się bardzo bliski 1. Błąd bezwzględny może rosnąć do dowolnie dużej wartości, a jednak błąd procentowy maleje do zera.

Jeśli zastanawiacie się, jak to możliwe, rozważmy przykład jakiejś hipotetycznej, skomplikowanej własności liczb pierwszych, którą można przybliżyć za pomocą ciągu liczb będących potęgami 100:

100, 10 000, 1 000 000, 100 000 000,

podczas gdy jej dokładne wartości są następujące:

101, 10 010, 1 000 100, 100 001 000,

czyli, jak widać, dodatkowa jedynka przesuwa się za każdym razem o jedno miejsce w lewo. W takiej sytuacji stosunki odpowiadających sobie liczb zbliżają się do wartości 1, choć różnice między nimi rosną:

1, 10, 100, 1000

i mogą się stać dowolnie duże. Z tego rodzaju zachowaniem mamy do czynienia wtedy, gdy błędy – czyli różnice między przybliżeniami i wartościami dokładnymi – rosną bez ograniczeń, ale wolniej od samych wartości.

Poszukiwania wzorów asymptotycznych związanych z liczbami pierwszymi przyczyniły się do rozwoju nowych metod z zakresu teorii liczb opierających się jednak nie na liczbach całkowitych, ale na analizie zespolonej. Analiza matematyczna jest ścisłym sformułowaniem rachunku różniczkowego i całkowego, który ma

dwa kluczowe aspekty. Pierwszy – rachunek różniczkowy – służy do badania, jak szybko jedna wielkość, nazywana funkcją, zmienia się względem drugiej. Na przykład położenie ciała zależy od – czyli jest funkcją – czasu, a szybkość zmian położenia ciała w miarę upływu czasu jest jego prędkością chwilową. Drugi aspekt – rachunek całkowy – pozwala na obliczanie powierzchni, objętości i podobnych wielkości za pomocą dodawania do siebie dużej liczby bardzo małych kawałków – proces ten nazywamy całkowaniem. Co ciekawe, okazuje się, że całkowanie jest odwrotnością różniczkowania. W swoim oryginalnym sformułowaniu, przedstawionym przez Isaaca Newtona i Gottfrieda Leibniza, rachunek różniczkowy i całkowy wymagał operowania nieskończenie małymi wielkościami, co dawało podstawy do kwestionowania logicznej poprawności takiej teorii. Ostatecznie te koncepcyjne nieścisłości udało się wyjaśnić dzięki wprowadzeniu pojęcia granicy, czyli wartości, do której możemy się zbliżyć na dowolnie małą odległość, choć nigdy nie zdołamy jej osiągnąć. W takim bardziej ścisłym ujęciu dziedzinę tę nazywamy analizą matematyczną.

W czasach Newtona i Leibniza rozważane wielkości były liczbami rzeczywistymi, dlatego również analiza matematyczna zajmuje się tego typu liczbami. Gdy jednak w matematyce zaczęto coraz powszechniej stosować liczby zespolone, dziedzina ta w zupełnie naturalny sposób rozszerzyła się na tego typu wielkości. Tak powstała analiza zespolona. Okazało się, że jest to niezwykle piękna i użyteczna dziedzina. Z punktu widzenia analizy matematycznej funkcje zespolone zachowują się dużo lepiej niż rzeczywiste. Oczywiście także one mają swoje dziwactwa, ale zalety działania na funkcjach zespolonych najczęściej przeważają nad wadami.

Matematycy z dużym zaskoczeniem odkryli, że własności arytmetyczne liczb całkowitych można z dużym powodzeniem zapisać w postaci funkcji zespolonych. Wcześniej w obydwu zbiorach liczbowych rozważano zupełnie różne problemy i posługiwano się różnymi metodami. Teraz jednak, dzięki analizie zespolonej dysponującej potężnym arsenałem metod, udało się odkryć szczególne cechy funkcji z dziedziny teorii liczb, co pozwoliło na

wyprowadzenie wzorów asymptotycznych i innych użytecznych formuł.

W 1859 roku niemiecki matematyk Bernhard Riemann powrócił do pewnej starej koncepcji sformułowanej kiedyś przez Eulera i nadał jej zupełnie nową postać tak zwanej funkcji dzeta. Jednym z wyników tych prac było otrzymanie *dokładnego* wzoru na liczbę liczb pierwszych mniejszych od ustalonej wartości granicznej. Jest to nieskończony szereg, ale specjaliści od analizy matematycznej przywykli już do takich wzorów. Nie chodzi tu tylko o jakąś sprytną, ale mało użyteczną sztuczkę – wzór ten pozwolił uczonym dowiedzieć się czegoś nowego na temat liczb pierwszych. Pojawił się tylko jeden mały szkopuł. Chociaż Riemann udowodnił, że jego wzór jest dokładny, najważniejszy wynikający z niego wniosek zależy od prawdziwości pewnego prostego stwierdzenia na temat funkcji dzeta – a Riemannowi nie udało się dowieść, że jest ono poprawne. Od tego czasu minęło już półtora wieku i wciąż nikomu nie udało się tego dokonać. Jest to tak zwana hipoteza Riemanna i jej udowodnienie stało się jednym z ambitnych celów matematyki teoretycznej.

W rozdziale 2 dowiedzieliśmy się, że w miarę jak przesuwamy się w kierunku coraz większych liczb, liczby pierwsze pojawiają się coraz rzadziej. Ponieważ matematycy wiedzieli, że nie możemy raczej liczyć na uzyskanie dokładnego wzoru na ich rozkład, postanowili poszukać zamiast tego jakichś statystycznych prawidłowości. W latach 1797–1798 Legendre wyliczył, ile jest liczb pierwszych mniejszych od różnych określonych wartości. Wykorzystał do tego tablice liczb pierwszych, które niedawno opublikowali Jurij Vega i Anton Felkel. Vega musiał wprost uwielbiać długie obliczenia: przygotował również tablice logarytmów, a w 1789 roku pobił rekord świata na najdokładniejsze wyznaczenie liczby π – obliczył ją z dokładnością do 140 miejsc po przecinku (z czego 126 było poprawnych). Podobne zamiłowanie do rachunków przejawiał Felkel. Jego główne dzieło wydane w 1776 roku to *Tafel aller einfachen Factoren der durch 2, 3, 5 nicht theilbaren Zahlen von 1 bis 10 000 000* (*Tablica wszystkich*

czynników pierwszych liczb do 10 milionów, z wyjątkiem liczb podzielnych przez 2, 3 lub 5). Istnieją proste sposoby na sprawdzenie czynników 2, 3 i 5 (wspomnieliśmy o nich w rozdziale 2), Felkel postanowił więc oszczędzić trochę miejsca, pomijając te liczby. Legendre odkrył doświadczalnie przybliżony wzór na liczbę liczb pierwszych mniejszych od danej wartości *x*, którą oznaczamy symbolem π(*x*). Jeśli dotychczas spotykaliście literę π jedynie w roli symbolu oznaczającego wartość 3,14159, to ten nowy zapis będzie początkowo wyglądał dość dziwnie, ale nietrudno się zorientować, jakie jest znaczenie tego symbolu w danym wypadku, nawet jeśli nie zwrócimy uwagi na inny krój czcionki. W książce Legendre'a z 1808 roku poświęconej teorii liczb znalazło się stwierdzenie, że π(*x*) wydaje się bardzo bliskie wartości $x/(\ln x - 1,08366)$.

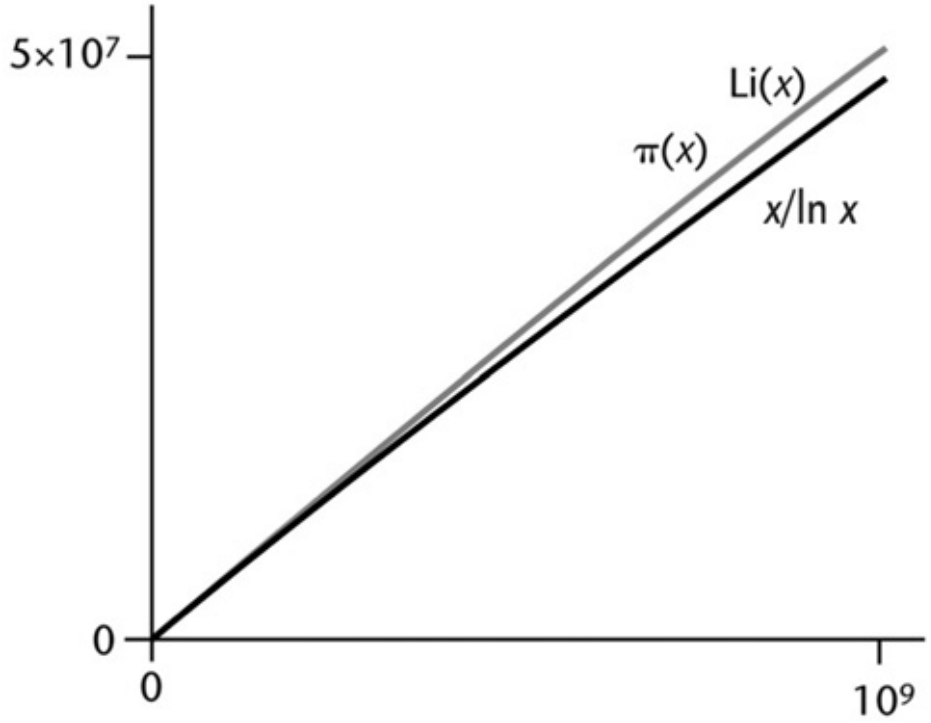
W napisanym w 1849 roku liście do astronoma Johanna Enckego Gauss wyznał, że gdy miał około piętnastu lat, zapisał w swoich tablicach logarytmicznych uwagę, iż liczba liczb pierwszych mniejszych lub równych wartości *x* wynosi $x/\ln x$ dla dużych *x*. Jak w wypadku wielu swoich odkryć, Gauss nie opublikował tego oszacowania, być może dlatego, że nie miał dowodu. W 1838 roku Dirichlet zwrócił jego uwagę na podobne, odkryte przez siebie przybliżenie. Sprowadza się ono do logarytmu całkowego[64]

$$\mathrm{Li}(x) = \int_0^x \frac{dt}{\ln t}$$

Stosunek Li(*x*) do $x/\ln x$ zmierza do 1, gdy *x* staje się duże, z czego wynika, że jeśli jeden z tych wzorów jest asymptotycznie równy π(*x*), to również drugi musi wykazywać taką cechę, jednak na podstawie ryciny 34 można dojść do (poprawnego) wniosku, że Li(*x*) jest lepszym przybliżeniem niż $x/\ln x$. Dokładność funkcji Li(*x*) robi duże wrażenie, na przykład:

π (1 000 000 000) = 50 847 534
Li (1 000 000 000) = 50 849 234,9.

Natomiast wzór $x/\ln x$ daje gorsze przybliżenie – w tym przykładzie jest to wartość 48 254 942,4.



Ryc. 34. W tej skali funkcje π(x) i Li(x) (zaznaczone szarym kolorem) są od siebie nieodróżnialne. Jednak wzór $x/\ln x$ (czarna prosta) daje wyraźnie mniejsze wartości. Na tym rysunku zmienna x jest na osi poziomej, natomiast wartości funkcji zaznaczono na osi pionowej.

Taki przybliżony wzór – wyrażony jako Li(*x*) lub $x/\ln x$ – stał się znany jako twierdzenie o liczbach pierwszych, przy czym słowa „twierdzenie” użyto tu w znaczeniu „hipoteza”. Udowodnienie, że wzory te są asymptotycznie równe π(*x*), stało się jednym z najważniejszych nierozwiązanych problemów teorii liczb. Wielu matematyków próbowało się z nim zmierzyć, wykorzystując do tego tradycyjne metody tej dziedziny, i niektórzy dotarli nawet dość blisko celu, jednak zawsze musieli w końcu przyjąć jakieś założenie, którego nie można było udowodnić. Najwyraźniej potrzebne były

jakieś nowe metody. Pojawiły się one za sprawą dziwnego sposobu zapisu dwóch starych, Euklidesowych twierdzeń o liczbach pierwszych.

Twierdzenie o liczbach pierwszych było odpowiedzią na twierdzenie Euklidesa głoszące, że liczby pierwsze ciągną się w nieskończoność. Z drugiego podstawowego twierdzenia Euklidesa wynika, że rozkład liczby na czynniki pierwsze jest jednoznaczny – każda dodatnia liczba całkowita jest iloczynem *jednego tylko* zestawu liczb pierwszych. W 1737 roku Euler uświadomił sobie, że pierwsze z tych twierdzeń można zapisać na gruncie analizy matematycznej w postaci dosyć dziwnego wzoru, a wtedy drugie twierdzenie stanie się wynikającym z niego prostym wnioskiem. Zacznijmy od przytoczenia tego wzoru, a potem spróbujmy go zrozumieć. Oto on:

$$\frac{1}{1-2^{-s}} \times \frac{1}{1-3^{-s}} \times \cdots \times \frac{1}{1-p^{-s}} \times \cdots$$

$$= \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \cdots$$

gdzie *p* przebiega przez wszystkie liczby pierwsze, a *s* jest stałą. Eulera interesował głównie przypadek, gdy *s* jest liczbą naturalną, ale jego wzór działa również dla wszystkich liczb rzeczywistych, pod warunkiem że *s* jest większe od 1. Warunek ten jest konieczny do tego, żeby prawa strona równości była zbieżna, to znaczy żeby miała sensowną wartość, gdy szereg ciągnie się w nieskończoność.

To niezwykły wzór. Po lewej stronie mamy iloczyn nieskończonej liczby wyrażeń zależnych wyłącznie od liczb pierwszych, po prawej zaś dodajemy do siebie nieskończenie wiele wyrażeń zależnych od wszystkich dodatnich liczb całkowitych. Wzór ten wyraża w języku analitycznym pewien związek między liczbami całkowitymi i liczbami pierwszymi. Najważniejszym związkiem tego rodzaju jest

jednoznaczność rozkładu na czynniki pierwsze i to właśnie on jest podstawą tego wzoru.

Omówmy teraz w zarysie główny krok potrzebny do pokazania, że za tym wszystkim kryje się sensowna idea. Korzystając z podstawowych reguł algebry, możemy przekształcić wyrażenie z *p* w szereg przypominający ten po prawej stronie naszej równości, ale wykorzystujący wyłącznie potęgi *p*. A zatem

$$\frac{1}{1-p^{-s}} = \frac{1}{1^s} + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \cdots$$

Po wymnożeniu ze sobą wszystkich takich szeregów dla wszystkich liczb pierwszych *p* i „rozpisaniu" ich tak, by uzyskać sumę prostych wyrazów, otrzymamy każdą kombinację potęg liczb pierwszych – to znaczy każdą liczbę całkowitą. Każda z nich występuje jako odwrotność swojej *s*-tej potęgi i pojawia się tylko raz, co wynika z jednoznaczności rozkładu na czynniki pierwsze. W ten sposób otrzymujemy więc szereg pojawiający się po prawej stronie naszego wzoru.

Nikomu nie udało się jeszcze znaleźć prostego algebraicznego wzoru na sumę tego szeregu, choć istnieje wiele wzorów wykorzystujących całki. Oznaczmy go więc specjalnym symbolem, grecką literą dzeta (ζ), i zdefiniujmy nową funkcję:

$$\zeta(s) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \cdots$$

Euler nie wprowadził tak naprawdę symbolu ζ i rozważał wyłącznie dodatnie wartości *s*, ale mimo to nazwijmy powyższy szereg funkcją dzeta Eulera. W oparciu o ten wzór Euler doszedł do wniosku, że jeśli pozwolimy, by *s* było coraz bliższe 1, to możemy uzyskać nieskończenie wiele liczb pierwszych. Gdyby liczb pierwszych nie było nieskończenie wiele, to wartość po lewej stronie wzoru musiałaby być skończona, natomiast szereg po drugiej stronie znaku

równości byłby nieskończony. Jest to sprzeczność, a zatem musi istnieć nieskończenie wiele liczb pierwszych. Głównym celem Eulera było otrzymanie wzorów takich jak $\zeta(2) = \pi^2/6$, określających sumę szeregu dla parzystych liczb całkowitych *s*. Nie wykorzystał swojego przełomowego pomysłu do przeprowadzenia bardziej zaawansowanych analiz.

Inni matematycy zauważyli to, co umknęło uwadze Eulera, i rozważyli przypadki, gdy *s* nie jest liczbą całkowitą. W dwóch artykułach z lat 1848 i 1850 rosyjski matematyk Pafnutij Czebyszew[65] przedstawił genialny pomysł: spróbujmy udowodnić twierdzenie o liczbach pierwszych, wykorzystując do tego analizę matematyczną. Wyszedł od związku między liczbami pierwszymi i analizą matematyczną wynikającego z funkcji dzeta Eulera. Nie udało mu się do końca osiągnąć postawionego przed sobą celu, ponieważ przyjął, że *s* jest liczbą rzeczywistą, i metody dostępne w tym zakresie analizy matematycznej okazały się zbyt ograniczone. Zdołał jednak dowieść, że gdy *x* jest duże, to stosunek π(*x*) do $x/\ln x$ zawiera się w przedziale między dwiema stałymi: jedna z nich jest nieco większa, a druga nieco mniejsza od 1. Oznaczało to prawdziwy postęp, nawet jeśli wynik nie był doskonały, ponieważ dzięki temu Czebyszew mógł udowodnić hipotezę Bertranda wysuniętą w 1845 roku, która głosi, że w przedziale między dowolną liczbą całkowitą i jej dwukrotnością istnieje przynajmniej jedna liczba pierwsza.

Wszystko już było gotowe na pojawienie się Riemanna. Również on zauważył, że funkcja dzeta jest kluczem do tajemnicy twierdzenia o liczbach pierwszych, ale by takie podejście miało szanse powodzenia, zaproponował znacznie ambitniejsze rozszerzenie – postanowił zdefiniować funkcję dzeta nie tylko dla liczb rzeczywistych, ale również dla zespolonych. Dobrym punktem wyjścia do tego jest szereg Eulera. Jest on zbieżny dla wszystkich rzeczywistych wartości *s* większych od 1 i okazuje się, że jeśli wykorzystamy taki sam wzór dla zespolonych wartości *s*, to szereg będzie zbieżny zawsze, gdy część rzeczywista liczby *s* jest większa

od 1. Riemann odkrył jednak, że można pójść jeszcze dalej. Posługując się metodą, którą nazwał przedłużeniem analitycznym, rozszerzył definicję funkcji $\zeta(s)$ na *wszystkie* liczby zespolone różne od 1. Ta wartość *s* została wykluczona, ponieważ dla *s* = 1 funkcja staje się nieskończona[66].

W 1859 roku Riemann spisał swoje przemyślenia na temat funkcji dzeta w artykule, którego tytuł w tłumaczeniu brzmi: „O liczbie liczb pierwszych mniejszych od danej wielkości”[67]. Podał też w nim jawny, dokładny wzór[68] na $\pi(x)$. Aby pokazać, jak pojawiają się miejsca zerowe funkcji dzeta, opiszmy tu prostszy wzór równoważny temu, jaki zaproponował Riemann w swoim artykule. Cały pomysł polega na policzeniu, ile jest liczb pierwszych lub ich potęg mniejszych od dowolnie wybranej wartości. Zamiast jednak po prostu zliczać każde pojawienie się tych liczb – tak jak zdefiniowano funkcję $\pi(x)$ dla liczb pierwszych – większym liczbom pierwszym nadamy większą wagę. Mówiąc dokładniej, wystąpienie dowolnej potęgi liczby pierwszej uwzględnimy z wagą równą logarytmowi naturalnemu z tej liczby. Na przykład w przedziale do 12 mamy następujące liczby pierwsze i ich potęgi:

$$2,\ 3,\ 4 = 2^2,\ 5,\ 7,\ 8 = 2^3,\ 9 = 3^2,\ 11,$$

zatem zliczając każde wystąpienie z tak zdefiniowaną wagą, mamy:

$$\ln 2 + \ln 3 + \ln 2 + \ln 5 + \ln 7 + \ln 2 + \ln 3 + \ln 11$$

co daje w sumie około 10,23.

Wykorzystując analizę matematyczną, możemy przekształcić wnioski wynikające z tego bardziej zawiłego sposobu zliczania liczb pierwszych w informację o tradycyjnym sposobie ich liczenia. Jednak ten nowy sposób pozwala wyprowadzić proste wzory, wydaje się więc, że wprowadzenie logarytmów jest niewielką ceną, którą warto zapłacić. Z dokładnego wzoru Riemanna wynika, że wykonując takie

zliczanie do określonej wartości *x* z uwzględnieniem wprowadzonych wag, otrzymamy wartość:

$$-\sum_{\rho}\frac{x^{\rho}}{\rho}+x-\frac{1}{2}\ln\left(1-x^{-2}\right)-\ln 2\pi$$

gdzie Σ oznacza sumę po wszystkich liczbach ρ, dla których $\zeta(\rho)$ jest równe zeru, z wyjątkiem ujemnych liczb parzystych. Są to tak zwane nietrywialne miejsca zerowe funkcji dzeta. Jej trywialnymi miejscami zerowymi są ujemne liczby parzyste –2, –4, –6, … Dla tych liczb funkcja dzeta przyjmuje wartość zero, bo tak wynika ze wzoru użytego w definicji analitycznego przedłużenia, ale okazuje się, że te miejsca zerowe nie są istotne we wzorze Riemanna ani w żadnym innym zastosowaniu.

Wzór ten może sprawiać wrażenie zawiłego, powiedzmy więc, co jest w nim najważniejsze: wymyślny sposób zliczania liczb pierwszych mniejszych od określonej wartości granicznej *x*, który za pomocą kilku analitycznych sztuczek można przekształcić w tradycyjny sposób liczenia liczb pierwszych, daje wynik *dokładnie* równy sumie prostego wyrażenia x^{ρ}/ρ po wszystkich nietrywialnych miejscach zerowych funkcji dzeta, powiększonej o wartość *nieskomplikowanej funkcji zmiennej x*. Każdy, kto zajmuje się analizą matematyczną, od razu dostrzeże, że udowodnienie twierdzenia o liczbach pierwszych jest równoważne dowiedzeniu, iż takie zliczanie do ustalonej granicy *x* z uwzględnieniem wag jest asymptotycznie równe *x*. Wykorzystując analizę zespoloną, można pokazać, że stwierdzenie to jest prawdziwe wtedy, gdy wszystkie nietrywialne miejsca zerowe funkcji dzeta są liczbami zespolonymi o części rzeczywistej zawierającej się w przedziale między 0 a 1. Czebyszewowi nie udało się tego udowodnić, ale był na tyle blisko, że uzyskał ważne informacje.

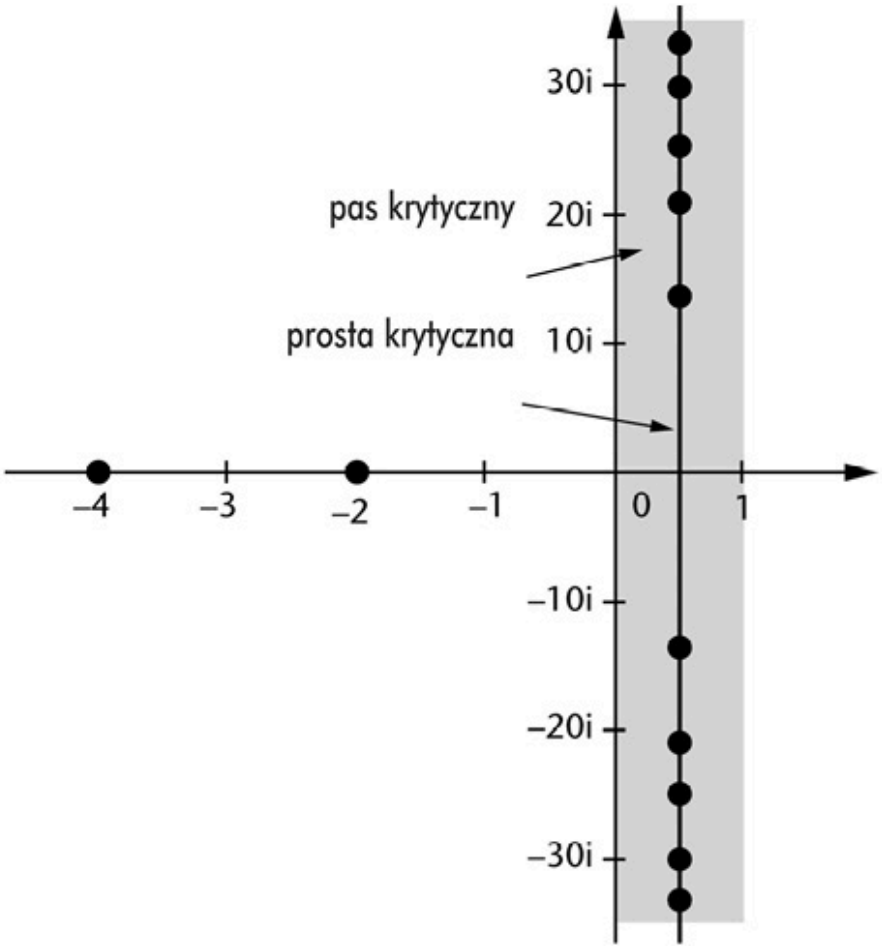
Dlaczego miejsca zerowe funkcji dzeta są tak ważne? Podstawowe twierdzenie analizy zespolonej głosi, że w pewnych ustalonych warunkach funkcję zmiennej zespolonej można w pełni określić,

podając wartości, w których jest ona równa zero lub nieskończona, wraz z dodatkowymi informacjami opisującymi jej zachowanie w tych punktach. Te szczególne miejsca nazywamy miejscami zerowymi i biegunami funkcji. Takie twierdzenie nie obowiązuje na gruncie analizy matematycznej operującej liczbami rzeczywistymi – to jeden z wielu powodów, dla których uczeni tak często korzystają z analizy zespolonej, mimo że wymaga ona posługiwania się pierwiastkiem kwadratowym z minus jeden. Funkcja dzeta ma tylko jeden biegun w punkcie $s = 1$, a więc całkowicie określają ją jej miejsca zerowe, choć nie wolno zapominać o tym pojedynczym biegunie.

Dla wygody Riemann najczęściej wykorzystywał w obliczeniach inną funkcję, którą oznaczał grecką literą ksi: $\xi(x)$. Funkcja ta jest blisko związana z funkcją dzeta i wynika z zastosowania metody analitycznego przedłużenia. Riemann zauważył, że

> Jest bardzo możliwe, iż wszystkie [miejsca zerowe funkcji ksi] są liczbami rzeczywistymi. Dobrze byłoby znaleźć ścisły dowód tego stwierdzenia, jednak po kilku wstępnych, bezowocnych próbach udowodnienia tej tezy odłożyłem na razie na bok dalsze prace w tym kierunku, ponieważ wydaje się, że nie jest to konieczne do osiągnięcia następnego celu moich badań.

To stwierdzenie na temat funkcji ksi można przekształcić w podobną tezę dotyczącą funkcji dzeta. Oznacza ono mianowicie, że wszystkie nietrywialne miejsca zerowe funkcji dzeta są liczbami zespolonymi postaci $\frac{1}{2}+it$, czyli innymi słowy, liczbami leżącymi na *prostej krytycznej* zawierającej wszystkie liczby o części rzeczywistej równej $\frac{1}{2}$ (zob. ryc. 35). Właśnie taka wersja jego uwagi jest słynną hipotezą Riemanna.



Ryc. 35. Miejsca zerowe funkcji dzeta, prosta krytyczna i pas krytyczny

Uwaga Riemanna sprawia wrażenie rzuconej mimochodem, jak gdyby jego hipoteza nie była zbyt ważna. Dla jego badań, których celem było udowodnienie twierdzenia o liczbach pierwszych, rzeczywiście nie ma ona znaczenia. Jednak w wielu innych wypadkach już tak nie jest. Prawdę mówiąc, hipoteza Riemanna jest powszechnie uważana za najważniejsze w całej matematyce pytanie bez odpowiedzi.

Aby zrozumieć dlaczego, musimy pójść śladem Riemanna jeszcze trochę dalej. Riemann miał pewne pomysły, jak sobie poradzić z twierdzeniem o liczbach pierwszych. Jego dokładny wzór pokazuje, jak trzeba się do tego zabrać: należy zrozumieć miejsca zerowe

funkcji dzeta lub, co równoważne, funkcji ksi. Nie jest do tego potrzebna pełna hipoteza Riemanna – wystarczy udowodnić, że wszystkie nietrywialne miejsca zerowe funkcji dzeta są liczbami zespolonymi, których część całkowita zawiera się w przedziale między 0 a 1. Innymi słowy, trzeba pokazać, że liczby te leżą w odległości $\frac{1}{2}$ od prostej krytycznej Riemanna, w tak zwanym pasie krytycznym. W takim razie suma po wszystkich miejscach zerowych funkcji dzeta, zapisana w przytoczonym wcześniej dokładnym wzorze Riemanna, jest stałą o skończonej wartości. W przypadku granicznym, gdy *x* staje się bardzo duże, mogłoby jej tam nawet w ogóle nie być. Spośród wszystkich wyrazów zapisanych w tym wzorze jedynym, który pozostaje ważny, gdy *x* staje się bardzo duże, jest samo *x*. Zatem zliczanie liczb pierwszych z uwzględnieniem wag jest asymptotycznie równe *x*, a to dowodzi prawdziwości twierdzenia o liczbach pierwszych. Jak na ironię, zadaniem miejsc zerowych funkcji dzeta jest więc udowodnienie, że nie wnoszą one istotnego wkładu do dokładnego wzoru.

Riemann nigdy nie doprowadził tych badań do końca. Nawet nic już więcej na ten temat nie napisał. Jednak wyzwanie to podjęli dwaj inni matematycy i pokazali, że intuicja go nie zawiodła. W 1896 roku Jacques Hadamard i Charles-Jean de la Vallée Poussin niezależnie potwierdzili poprawność twierdzenia o liczbach pierwszych, dowodząc, że wszystkie nietrywialne miejsca zerowe funkcji dzeta leżą w pasie krytycznym. Przedstawione przez nich dowody są bardzo skomplikowane i wymagają głębokiej wiedzy, ale okazały się poprawne. W ten sposób powstał nowy, bardzo ważny dział matematyki – analityczna teoria liczb. Dziedzina ta znalazła zastosowanie w całej teorii liczb i dzięki niej udało się rozwiązać stare problemy, a także odkryć nowe prawidłowości. Inni matematycy znaleźli później prostsze analityczne dowody twierdzenia o liczbach pierwszych, a Atle Selberg i Paul Erdős odkryli nawet bardzo skomplikowany dowód, który w ogóle nie wymaga użycia analizy zespolonej. Tymczasem jednak ideę Riemanna zastosowano już do udowodnienia mnóstwa ważnych twierdzeń, w tym przybliżeń wielu funkcji z dziedziny teorii liczb, zatem ich nowy dowód był tylko

ciekawostką pozbawioną większego znaczenia. W 1980 roku Donald Newman znalazł znacznie prostszy dowód wykorzystujący tylko jedno z najbardziej podstawowych twierdzeń analizy zespolonej, a mianowicie twierdzenie Cauchy'ego.

Chociaż Riemann stwierdził, że udowodnienie jego hipotezy nie jest konieczne, by dotrzeć do celu, jaki przed sobą postawił, to okazało się, że hipoteza ta ma kluczowe znaczenie dla wielu innych pytań z zakresu teorii liczb. Zanim omówimy jednak hipotezę Riemanna, przyjrzyjmy się niektórym twierdzeniom, których prawdziwość byłaby pewna, gdyby udało się ją udowodnić.

Jeden z najważniejszych wniosków wynikających z tej hipotezy dotyczy rozmiaru błędu w twierdzeniu o liczbach pierwszych. Twierdzenie to głosi, że dla dużych x iloraz funkcji $\pi(x)$ i $\mathrm{Li}(x)$ zbliża się coraz bardziej do 1. Innymi słowy, wielkość różnicy między tymi funkcjami, *w odniesieniu do wielkości x*, kurczy się do zera. Jednak faktyczna różnica może cały czas się powiększać (i rzeczywiście tak jest). Rośnie jedynie wolniej[69] niż sama wartość x. Z doświadczeń komputerowych wynika, że wielkość tego błędu jest w przybliżeniu proporcjonalna do $\sqrt{x}\ln x$. Stwierdzenie to daje się udowodnić przy założeniu, że hipoteza Riemanna jest prawdziwa. W 1901 roku Helge von Koch pokazał, że hipoteza Riemanna jest logicznie równoważna oszacowaniu

$$\left|\pi(x)-\mathrm{Li}(x)\right| \le \frac{1}{8\pi}\sqrt{x}\ln x$$

dla wszystkich $x \ge 2657$. Symbol „||" oznacza w tym wzorze wartość bezwzględną, czyli to, że zapisaną między pionowymi kreskami różnicę funkcji należy pomnożyć przez ±1, tak by była dodatnia. Wzór ten jest najlepszym możliwym ograniczeniem wartości różnicy funkcji $\pi(x)$ i $\mathrm{Li}(x)$.

Z hipotezy Riemanna wynika jeszcze wiele innych oszacowań różnych funkcji z dziedziny teorii liczb. Na przykład jej poprawność

oznacza, że suma dzielników liczby *n* jest mniejsza od

$$e^{\gamma} n \ln(\ln n)$$

dla wszystkich $n \geq 5040$, gdzie $\gamma = 0{,}57721\ldots$ jest stałą Eulera[70]. Może się wydawać, że to jakieś dziwactwa, ale dobre oszacowania ważnych funkcji są niezwykle istotne w wielu zastosowaniach i większość specjalistów od teorii liczb dałaby sobie uciąć prawą rękę za możliwość udowodnienia któregoś z nich.

Na podstawie hipotezy Riemanna możemy również określić, jak duże mogą być odstępy między kolejnymi liczbami pierwszymi. Średnią wielkość takiego odstępu możemy wyznaczyć z twierdzenia o liczbach pierwszych: średnio rzecz biorąc, odstęp między dowolną liczbą pierwszą *p* i liczbą pierwszą bezpośrednio po niej następującą jest porównywalny z $\ln p$. Niektóre odstępy są mniejsze, inne większe, jednak życie matematyków byłoby znacznie prostsze, gdyby wiedzieli, jak duży może być największy taki odstęp. W 1936 roku Harald Cramér udowodnił, że jeśli hipoteza Riemanna jest poprawna, to taki odstęp dla liczby pierwszej *p* jest nie większy niż stała razy $\sqrt{p} \ln p$.

Prawdziwe znaczenie hipotezy Riemanna jest jednak o wiele głębsze. Istnieją pewne daleko sięgające uogólnienia, a także uzasadnione przewidywania, z których wynika, że jeśli komuś uda się udowodnić poprawność hipotezy Riemanna, to osoba ta zdoła zapewne udowodnić również uogólnioną wersję tej hipotezy. To z kolei zapewni matematykom daleko idącą kontrolę nad różnorodnymi działami teorii liczb.

Uogólniona hipoteza Riemanna wynika z pewnego dokładniejszego opisu liczb pierwszych. Wszystkie liczby pierwsze większe od 2 są nieparzyste i w rozdziale 2 przekonaliśmy się, że można je podzielić na dwa rodzaje – pierwszą grupę stanowią liczby pierwsze o 1 większe od pewnej wielokrotności liczby 4, drugą zaś te, które są większe o 3 od pewnej wielokrotności liczby 4. Mówimy, że są to

liczby postaci $4k + 1$ lub $4k + 3$, gdzie k jest liczbą, przez którą należy pomnożyć 4, żeby je uzyskać. Oto krótka lista kilku początkowych liczb pierwszych każdego rodzaju wraz z odpowiadającymi im wielokrotnościami 4:

wielokrotność 4	0	4	8	12	16	20	24	28	32	36
plus 1	0	5	•	13	17	•	•	29	•	37
plus 3	3	7	11	•	19	23	•	•	•	•

Kropka oznacza, że dana liczba nie jest liczbą pierwszą.

Ile jest liczb pierwszych każdego rodzaju? Jak wygląda ich rozkład wśród liczb pierwszych lub wśród wszystkich liczb całkowitych? Bez większego wysiłku możemy zmodyfikować dowód Euklidesowy tego, że istnieje nieskończenie wiele liczb pierwszych, i udowodnić, że istnieje również nieskończenie wiele liczb pierwszych postaci $4k + 3$. Znacznie trudniej jest udowodnić, że również liczb pierwszych postaci $4k + 1$ jest nieskończenie wiele. Można tego dokonać, ale wymaga to wykorzystania dość trudnych twierdzeń. Różnica wynika z faktu, że każda liczba postaci $4k + 3$ ma jakiś czynnik takiej samej postaci, natomiast stwierdzenie takie nie zawsze jest już prawdziwe dla liczb postaci $4k + 1$.

W tych liczbach nie ma nic magicznego. Poza liczbami 2 i 3 wszystkie pozostałe liczby pierwsze są postaci albo $6k + 1$, albo $6k + 5$ – możemy więc w odniesieniu do nich postawić podobne pytania. Idąc dalej, można zauważyć, że wszystkie liczby pierwsze z wyjątkiem liczby 5 mają jedną z czterech postaci: $5k + 1$, $5k + 2$, $5k + 3$ lub $5k + 4$. Pomijamy tu $5k$, ponieważ są to wielokrotności 5, wszystkie więc, z wyjątkiem samej liczby 5, są liczbami złożonymi.

Bez trudu możemy wysunąć sensowne propozycje wszystkich tego rodzaju przypadków – liczb pierwszych występujących w ciągu arytmetycznym. Przypadek $5k$ jest typowy. Wystarczy przeprowadzić kilka doświadczeń, żeby szybko dojść do wniosku, że liczby czterech wymienionych przed chwilą typów mogą być liczbami pierwszymi mniej więcej z takim samym prawdopodobieństwem. Oto tablica dla tego przypadku:

wielokrotność 5	5	10	15	20	25	30	35	40
plus 1	•	11	•	•	•	31	•	41
plus 2	7	•	17	•	•	•	37	•
plus 3	•	13	•	23	•	•	•	43
plus 4	•	•	19	•	29	•	•	•

Zatem liczb każdego z tych rodzajów powinno być nieskończenie wiele i średnio rzecz biorąc, na każdą z tych postaci ma przypadać mniej więcej jedna czwarta liczb pierwszych mniejszych od określonej wartości.

W przypadku niektórych z takich postaci można prosto udowodnić, że liczb pierwszych danego rodzaju jest nieskończenie wiele, dla innych form zapisu liczb pierwszych dowód taki jest bardziej skomplikowany, ale do połowy XIX wieku nikomu nie udało się udowodnić, że istnieje nieskończenie wiele liczb pierwszych każdej możliwej postaci, nie mówiąc już o udowodnieniu tego, że proporcje takich liczb są mniej więcej równe. Gdy w 1785 roku Lagrange pracował nad swoim prawem wzajemności reszt kwadratowych – jest to ważna własność kwadratów liczb w arytmetyce z modułem będącym liczbą pierwszą – założył, że tak jest, nie przedstawiając na to dowodu. Nie ulega więc wątpliwości, że są to bardzo przydatne wnioski i ktoś powinien je wreszcie udowodnić. W 1837 roku Dirichlet odkrył, jak można wykorzystać pomysły Riemanna związane z twierdzeniem o liczbach pierwszych do udowodnienia obydwu tych tez. Pierwszy krok polega na zdefiniowaniu odpowiedników funkcji dzeta dla liczb pierwszych takiej postaci. Funkcje te nazywamy funkcjami *L* (lub *L*-funkcjami) Dirichleta. Jako przykład weźmy funkcję *L* dla $4k + 1$, $4k + 3$:

$$L(s,\chi) = 1 - 3^{-s} + 5^{-s} - 7^{-s} + 9^{-s} - \cdots$$

gdzie współczynniki mają wartość +1 dla liczb postaci $4k + 1$, –1 dla liczb $4k + 3$ i 0 dla pozostałych. Grecka litera χ jest tak zwanym charakterem Dirichleta, który przypomina nam o tym, że należy użyć takich właśnie znaków.

Dla funkcji dzeta Riemanna znaczenie ma nie tylko szereg, ale również jego analityczne przedłużenie, które sprawia, że funkcja nabiera sensu dla wszystkich liczb zespolonych. To samo dotyczy funkcji *L* i Dirichlet zdefiniował odpowiednie analityczne przedłużenie. Dzięki wykorzystaniu pomysłów użytych w dowodzie twierdzenia o liczbach pierwszych udało mu się udowodnić analogiczne twierdzenie dla liczb pierwszych określonej postaci. Na przykład liczba liczb pierwszych postaci $5k + 1$ mniejszych lub równych x jest asymptotycznie równa $\mathrm{Li}(x)/4$ i tak samo jest w pozostałych trzech przypadkach $5k + 2$, $5k + 3$ oraz $5k + 4$. W szczególności istnieje nieskończenie wiele liczb pierwszych każdej z tych postaci.

Funkcja dzeta Riemanna jest szczególnym przypadkiem funkcji *L* Dirichleta dla liczb pierwszych postaci $1k + 0$, a więc dla wszystkich liczb pierwszych. Uogólniona hipoteza Riemanna stanowi oczywiste rozszerzenie jego pierwotnej tezy: miejsca zerowe dowolnej funkcji *L* Dirichleta są liczbami zespolonymi z częścią rzeczywistą równą $\frac{1}{2}$ lub „trywialnymi miejscami zerowymi" o części rzeczywistej albo ujemnej, albo większej od 1.

Jeśli uogólniona hipoteza Riemanna jest prawdziwa, to również pierwotna jej wersja musi być prawdziwa. Wiele wniosków wynikających z uogólnionej hipotezy Riemanna jest podobnych do tych, jakie udało się wysnuć z pierwotnej jej wersji. Można na przykład wyprowadzić podobne ograniczenia na wielkość błędu dla odpowiednich wersji twierdzenia o liczbach pierwszych odnoszących się do liczb dowolnej określonej postaci. Z uogólnionej hipotezy Riemanna wynika jednak również wiele wniosków, które różnią się całkowicie od tego, co można stwierdzić na podstawie zwykłej hipotezy Riemanna. W 1917 roku Godfrey Harold Hardy i John Edensor Littlewood dowiedli, że z uogólnionej hipotezy Riemanna wynika hipoteza Czebyszewa, co oznacza, iż liczby pierwsze postaci $4k + 3$ są częstsze (w pewnym ściśle określonym sensie) od liczb pierwszych postaci $4k + 1$. Z twierdzenia Dirichleta wynika, że w szerszej perspektywie oba rodzaje są jednakowo prawdopodobne,

ale to nie przeszkadza, by liczby pierwsze postaci $4k + 3$ wygrywały z tymi postaci $4k + 1$, jeśli przyjmiemy odpowiednie zasady gry.

Z uogólnionej hipotezy Riemanna wypływają również ważne wnioski dla testów pierwszości, takich jak test Millera z 1976 roku, o którym wspomnieliśmy w rozdziale 2. Jeśli uogólniona hipoteza Riemanna jest prawdziwa, oznacza to, że test Millera jest wydajnym algorytmem. Również oszacowania wydajności innych testów zależą od uogólnionej hipotezy Riemanna. Istnieją też ważne zastosowania w algebraicznej teorii liczb. Jak pamiętamy z rozdziału 7, Dedekind sformułował na nowo idealne liczby Kummera i wprowadził nowe, podstawowe pojęcie ideału. Rozkład na czynniki pierwsze w pierścieniach liczb algebraicznych całkowitych jest możliwy, ale nie musi być jednoznaczny. Rozkład na czynniki pierwsze ideałów jest znacznie bardziej uporządkowany: tu prawdą jest istnienie takiego rozkładu i jego jednoznaczność. Można więc przenieść wszystkie pytania na temat czynników pierwszych na grunt ideałów. W szczególności istnieje pojęcie „ideału pierwszego", który jest sensownym, łatwym do wyznaczenia odpowiednikiem liczby pierwszej.

Wiedząc to wszystko, wypada zapytać, czy znaleziony przez Eulera związek między zwyczajnymi liczbami pierwszymi a funkcją dzeta ma jakiś swój odpowiednik w przypadku ideałów pierwszych. Jeśli tak, to cały potężny arsenał metod analitycznej teorii liczb będzie można zastosować do liczb algebraicznych. Okazuje się, że można tego dokonać i wynikają z tego głębokie i niezwykle ważne konsekwencje. Prowadzi to do uzyskania funkcji dzeta Dedekinda – po jednej dla każdego zbioru liczb algebraicznych. Istnieje głęboki związek między własnościami funkcji dzeta Dedekinda wynikającymi z analizy zespolonej a arytmetyką liczb pierwszych odpowiednich liczb algebraicznych całkowitych. No i oczywiście istnieje też odpowiednik hipotezy Riemanna: wszystkie nietrywialne miejsca zerowe funkcji dzeta Dedekinda leżą na prostej krytycznej. Określenie „uogólniona hipoteza Riemanna" obejmuje obecnie również taką tezę.

Nawet jednak to uogólnienie nie jest jeszcze wszystkim, co można powiedzieć o funkcji dzeta. Zainspirowała ona matematyków do zdefiniowania analogicznych funkcji w kilku innych gałęziach,

od algebry abstrakcyjnej po teorię układów dynamicznych. We wszystkich tych dziedzinach istnieją jeszcze głębiej sięgające odpowiedniki hipotezy Riemanna. W niektórych wypadkach udało się wykazać ich prawdziwość. W 1974 roku Pierre Deligne udowodnił taką analogiczną hipotezę dla różnych ciał skończonych. W przypadku uogólnień znanych jako funkcje dzeta Selberga odpowiednik hipotezy Riemanna jest prawdziwy. Podobnie rzecz ma się z funkcją dzeta Gossa. Istnieją jednak również inne uogólnienia, na przykład funkcje dzeta Epsteina, dla których odpowiednia wersja hipotezy Riemanna okazuje się nieprawdziwa. W przypadku funkcji dzeta Epsteina na prostej krytycznej leży nieskończenie wiele nietrywialnych miejsc zerowych, ale niektóre znajdują się jednak poza tą prostą, co udowodnił Edward Titchmarsh w 1986 roku. Równocześnie dla tych funkcji dzeta nie ma odpowiednika eulerowskiego wzoru iloczynowego, zatem nie przypominają one funkcji dzeta Riemanna w pewnym aspekcie, który może okazać się kluczowy.

Badaczom udało się zebrać imponującą liczbę dowodów poszlakowych przemawiających za prawdziwością hipotezy Riemanna – zarówno w jej postaci oryginalnej, jak i w wersjach uogólnionych. Przy założeniu prawdziwości tej hipotezy można wyprowadzić wiele wspaniałych wniosków. Żadnego z nich nie udało się jeszcze nikomu obalić – dokonanie tego byłoby dowodem na nieprawdziwość hipotezy Riemanna, ale nie potrafimy dowieść ani jej poprawności, ani niepoprawności. W społeczności matematyków panuje powszechne przeświadczenie, że udowodnienie oryginalnej hipotezy Riemanna utorowałoby drogę do udowodnienia również jej uogólnień. Lepsze wyniki może jednak przynieść zmierzenie się z uogólnioną wersją w całej jej okazałości, z wykorzystaniem olbrzymiego bogactwa dostępnych obecnie metod, i dopiero w drugiej kolejności wyprowadzenie oryginalnej hipotezy jako szczególnego przypadku.

Mamy również olbrzymią liczbę dowodów doświadczalnych potwierdzających prawdziwość hipotezy Riemanna, a przynajmniej

coś, co sprawia wrażenie olbrzymiej liczby tego typu faktów – to jednak może szybko się zmienić, jeśli ktoś wyleje nam na głowę kubeł zimnej wody. Carl Ludwig Siegel sądzi, że Riemann wyznaczył numerycznie kilka pierwszych miejsc zerowych swojej funkcji dzeta, ale nie opublikował tych wyników. Są to następujące liczby zespolone:

$$\frac{1}{2} \pm 14{,}135i$$

$$\frac{1}{2} \pm 21{,}022i$$

$$\frac{1}{2} \pm 25{,}011i$$

Nietrywialne miejsca zerowe zawsze występują w tego typu parach z symbolem „±”. Napisaliśmy tu $\frac{1}{2}$, a nie 0,5, ponieważ w tych przypadkach wartość części rzeczywistej udało się wyznaczyć *dokładnie* dzięki wykorzystaniu ogólnych wyników analizy zespolonej i znanych własności funkcji dzeta. To samo dotyczy omawianych poniżej rezultatów obliczeń komputerowych. Nie wynika z nich jedynie, że miejsca zerowe są położone bardzo blisko prostej krytycznej – pokazują one, że punkty te leżą dokładnie na tej prostej.

W 1903 roku Jørgen Gram dowiódł numerycznie, że pierwsze dziesięć miejsc zerowych (a właściwie ich par postaci ±) leży na prostej krytycznej. Do 1935 roku Titchmarsh zwiększył tę liczbę do 195. W 1936 roku Titchmarsh i Leslie Comrie wykazali, że pierwsze 1041 par miejsc zerowych leży na prostej krytycznej – był to ostatni przypadek przeprowadzenia tego typu obliczeń ręcznie. Alan Turing jest najbardziej znany dzięki pracom, które prowadził w czasie wojny w Bletchley Park, gdzie pomagał w złamaniu kodu Enigmy, oraz temu, że położył podwaliny pod współczesną informatykę i badania sztucznej inteligencji. Interesował się jednak również analityczną

teorią liczb. W 1953 roku odkrył bardziej wydajną metodę wyznaczania miejsc zerowych funkcji dzeta i wykorzystując do tego komputer, wywnioskował, że pierwsze 1104 pary miejsc zerowych leżą na prostej krytycznej. Później pojawiły się kolejne dowody, że wszystkie miejsca zerowe mniejsze od określonej wartości granicznej leżą na prostej krytycznej. Obecny rekord z 2004 roku, należący do Yannicka Saoutera i Patricka Demichela, wynosi 10 bilionów (10^{13}). Różni matematycy i informatycy badali także inne przedziały miejsc zerowych. Na razie wszystkie nietrywialne miejsca zerowe, które udało się wyznaczyć, leżą na prostej krytycznej.

Mogłoby się wydawać, że to przesądza sprawę, ale matematycy odnoszą się z rezerwą do tego typu dowodów i mają ku temu dobre powody. Liczby takie jak 10 bilionów mogą się wydawać bardzo duże, ale w teorii liczb bardzo często istotny jest tak naprawdę logarytm z liczby, który jest proporcjonalny do liczby jej cyfr. Logarytm z 10 bilionów to niespełna 30. Prawdę mówiąc, rozwiązanie wielu problemów zależy od logarytmu z logarytmu, a nawet od logarytmu z logarytmu z logarytmu. W takiej sytuacji 10 bilionów to *bardzo mało* i zgromadzone dowody numeryczne dla wartości sięgających 10 bilionów mają naprawdę nieduże znaczenie.

Istnieją również pewne ogólne analityczne dowody, których tego typu zastrzeżenia nie dotyczą. Hardy i Littlewood udowodnili, że na prostej krytycznej leży nieskończenie wiele miejsc zerowych. Inni matematycy pokazali, że prawie wszystkie miejsca zerowe leżą bardzo blisko prostej krytycznej (określenie „prawie wszystkie” ma tu ściśle określone znaczenie). Selberg dowiódł, że niezerowy odsetek miejsc zerowych leży na prostej krytycznej. Norman Levinson pokazał, że odsetek ten to przynajmniej jedna trzecia, a obecnie wartość tę udało się zwiększyć do co najmniej 40 procent. Wszystkie te wyniki sugerują, że jeśli hipoteza Riemanna jest nieprawdziwa, to miejsca zerowe nieleżące na prostej krytycznej muszą być bardzo duże i należeć do rzadkości. Niestety, wynika z tego mało optymistyczny wniosek, że jeśli takie wyjątki istnieją, to znalezienie ich będzie niezwykle trudne.

Czy warto się tym przejmować? Przecież zgromadzone już dowody numeryczne powinny przekonać każdego sensownie myślącego człowieka, prawda? Niestety, nie. Nie przekonują one matematyków i nie wynika to wcale z ich przewrażliwienia – zachowują się, jak przystało na sensownie myślące osoby. W matematyce w ogólności, a w teorii liczb w szczególności, fakt zgromadzenia pozornie pokaźnego zbioru dowodów „doświadczalnych" ma często mniejsze znaczenie, niż sobie wyobrażamy.

Doskonałym tego potwierdzeniem jest przypadek hipotezy Pólyi, którą sformułował w 1919 roku węgierski matematyk George Pólya. Hipoteza ta głosi, że przynajmniej połowa wszystkich liczb naturalnych nieprzekraczających określonej wartości ma nieparzystą liczbę czynników pierwszych. W tym wypadku pomija się liczbę 1 i zaczyna od 2, a czynniki wielokrotne należy liczyć oddzielnie. W tabeli 2 pokazano, ile czynników pierwszych mają liczby od 2 do 20. W kolumnie „udział procentowy" zapisano wartość mówiącą, jaki odsetek liczb z przedziału do bieżącej wartości ma nieparzystą liczbę czynników pierwszych.

Wszystkie udziały procentowe w ostatniej kolumnie są większe od 50 procent i przeprowadzając szerzej zakrojone obliczenia, można dojść do wniosku, że tak sformułowana hipoteza powinna być prawdziwa. W 1919 roku, gdy nie było jeszcze komputerów, matematykom wykonującym tego typu doświadczenia nie udało się znaleźć żadnej liczby, która przeczyłaby prawdziwości tej tezy. Jednak w 1958 roku Brian Haselgrove posłużył się analityczną teorią liczb do udowodnienia, że ta hipoteza musi być nieprawdziwa dla pewnej liczby – dokładnie rzecz biorąc, dla wartości mniejszej niż $1{,}845 \times 10^{361}$. Po pojawieniu się komputerów Sherman Lehman dowiódł, że hipoteza ta nie jest spełniona dla liczby 906 180 359. W 1980 roku Minoru Tanaka udowodnił, że najmniejszym tego typu kontrprzykładem jest 906 150 257. Matematycy mogliby więc zgromadzić olbrzymią liczbę dowodów potwierdzających prawdziwość tej hipotezy dla liczb sięgających niemal miliarda, mimo że tak naprawdę jest ona fałszywa.

Liczba	Rozkład	Liczba czynników pierwszych	Udział procentowy
2	2	1	100
3	3	1	100
4	2^2	2	66
5	5	1	75
6	2×3	2	60
7	7	1	66
8	2^3	3	71
9	3^2	2	62
10	2×5	2	55
11	11	1	60
12	$22^2 \times 3$	3	63
13	13	1	66
14	2×7	2	61
15	3×5	2	57
16	2^4	4	53
17	17	1	56
18	2×3^2	3	59
19	19	1	61
20	$2^2 \times 5$	3	63

Tabela 2. Udział procentowy liczb nieprzekraczających określonej wartości, które mają nieparzystą liczbę czynników pierwszych

Mimo wszystko dobrze wiedzieć, że liczba 906 150 257 jest niezwykle interesująca.

Oczywiście obecnie komputery umożliwiłyby obalenie tej hipotezy w ciągu kilku sekund – wystarczyłoby jedynie napisać odpowiedni program. Czasami jednak nawet komputery nie są w stanie nam pomóc. Klasycznym tego przykładem jest liczba Skewesa, w której przypadku nagromadzono olbrzymią liczbę dowodów numerycznych przemawiających za poprawnością pewnej słynnej hipotezy, chociaż ostatecznie okazało się, że jest ona fałszywa. Ta gigantyczna liczba pojawiła się przy okazji badań problemu blisko związanego z hipotezą Riemanna – przybliżenia funkcji $\pi(x)$ za pomocą $\mathrm{Li}(x)$. Jak

już wiemy, twierdzenie o liczbach pierwszych głosi, że iloraz tych dwóch wielkości zmierza do 1, gdy x staje się bardzo duże. Z obliczeń numerycznych wydaje się wynikać, że jedna z tych wielkości zawsze będzie górą, czyli że tak zdefiniowany iloraz będzie zawsze mniejszy od 1. Innymi słowy, funkcja $\pi(x)$ powinna być zawsze mniejsza od $\mathrm{Li}(x)$. W 2008 roku Tadej Kotnik przeprowadził obliczenia pokazujące, że jest to prawdą dla wszystkich wartości x mniejszych od 10^{14}. Do 2012 roku Douglas Stoll i Patrick Demichel przesunęli tę granicę do wartości 10^{18} i taką samą liczbę otrzymał niezależnie od nich Andry Kulsha. Wyniki uzyskane przez Tomása Oliveirę e Silvę sugerują, że granicę tę można zwiększyć do 10^{20}.

Mogłoby się wydawać, że to przesądza sprawę. To silniejsze dowody od najlepszych wyników uzyskanych metodami numerycznymi dla hipotezy Riemanna. Jednak w 1914 roku Littlewood udowodnił, że hipoteza ta jest nieprawdziwa – i to zdecydowanie. Gdy x przechodzi przez wszystkie dodatnie liczby rzeczywiste, różnica $\pi(x) - \mathrm{Li}(x)$ zmienia znak (z ujemnego na dodatni lub odwrotnie) *nieskończenie często*. W szczególności funkcja $\pi(x)$ jest *większa* od $\mathrm{Li}(x)$ dla pewnych odpowiednio dużych wartości x. Z dowodu Littlewooda nie wynika jednak żadna wskazówka sugerująca, jak duża jest to wartość.

W 1933 roku jego student ze Związku Południowej Afryki, matematyk Stanley Skewes, oszacował wymaganą wielkość wartości x – ustalił, że nie powinna ona przekraczać $10\hat{}10\hat{}10\hat{}34$, gdzie symbol „ˆ” oznacza podniesienie do potęgi. Jest to tak gigantyczna liczba, że gdyby ktoś zechciał wydrukować wszystkie jej cyfry w książce – choć byłaby to dość nudna lektura: na kartach książki widniałaby cyfra 1 i nieskończenie długi ciąg zer – to okazałoby się, że Wszechświat nie jest wystarczająco duży, by ją pomieścić, nawet gdyby każda cyfra miała rozmiar cząstki subatomowej. Dokonując tego oszacowania, Skewes musiał założyć prawdziwość hipotezy Riemanna. W 1955 roku znalazł sposób na uniknięcie odwoływania się do hipotezy Riemanna, ale przyszło mu za to zapłacić pewną cenę: wielkość jego oszacowania wzrosła do $10\hat{}10\hat{}10\hat{}963$.

Są to tak duże liczby, że ich wielkości nie oddaje nawet przymiotnik „astronomiczne”, ale późniejsze badania pozwoliły je obniżyć do takiego poziomu, że można już było nazwać je „kosmologicznymi”. W 1966 roku Lehman zastąpił liczby Skewesa wartością 10^{1165}. W 1987 roku te Riele obniżył ją dalej do 7×10^{370}, a w 2000 roku Carter Bays i Richard Hudson zmniejszyli ją jeszcze bardziej, do $1{,}39822 \times 10^{316}$. Następnie Kuok Fai Chow i Roger Plymen zdołali jeszcze trochę z niej uskubnąć i zeszli do wartości $1{,}39801 \times 10^{316}$. Mogłoby się wydawać, że to znikoma poprawa, ale nowa wartość była około 2×10^{313} razy mniejsza. Saouter i Demichel dokonali kolejnego postępu, uzyskując $1{,}3971667 \times 10^{316}$. Tymczasem w 1941 roku Aurel Wintner udowodnił, że pewien niewielki, ale niezerowy odsetek liczb całkowitych spełnia nierówność $\pi(x) > \mathrm{Li}(x)$. W 2011 roku Stoll i Demichel obliczyli pierwsze 200 miliardów miejsc zerowych funkcji dzeta, co pozwoliło im zapanować nad funkcją $\pi(x)$ dla wartości x sięgających $10^{10\ 000\ 000\ 000\ 000}$. Dzięki temu znaleźli dowód[71] na to, że gdy x jest mniejsze od $3{,}17 \times 10^{114}$, to $\pi(x)$ jest mniejsze od $\mathrm{Li}(x)$. Zatem w tym konkretnym przypadku wszystkie dowody, które możemy uzyskać dla wartości sięgających przynajmniej do 10^{18}, a bardzo możliwe, że i do 10^{114}, a nawet dalej, są całkowicie mylące. Kapryśni bogowie teorii liczb nieźle się bawią kosztem śmiertelników.

Dokonano już wielu prób udowodnienia lub obalenia hipotezy Riemanna. Na swojej stronie internetowej zatytułowanej *Proposed proofs of the Riemann hypothesis* (*Propozycje dowodów hipotezy Riemanna*)[72] Matthew Watkins wymienia około 50 takich prób, które podjęto po 2000 roku. W wielu z nich znaleziono błędy i żadnej z nich eksperci nie uznali za poprawną.

Jedną z najgłośniejszych w ostatnich latach prób rozprawienia się z hipotezą Riemanna była praca Louisa de Brangesa z 2002 roku. De Branges rozesłał kolegom obszerny maszynopis, w którym twierdził,

że udowodnił hipotezę Riemanna dzięki zastosowaniu pewnej gałęzi analizy matematycznej zajmującej się operatorami na przestrzeniach o nieskończonej liczbie wymiarów, znanej jako analiza funkcjonalna. Były pewne podstawy, by potraktować de Brangesa poważnie. Wcześniej rozesłał dowód hipotezy Bieberbacha dotyczącej rozwinięć w szereg funkcji zespolonych. Chociaż jego pierwotny dowód zawierał błędy, ostatecznie okazało się, że zastosowane przez niego podejście przyniosło oczekiwany skutek. Obecnie jednak wydaje się, że mamy solidne podstawy, by uważać, iż zaproponowana przez de Brangesa metoda udowodnienia hipotezy Riemanna nie ma szans zadziałać. Brian Conrey i Xian-Jin Li natrafili na pewne przeszkody, których – wszystko na to wskazuje – nie da się pokonać[73].

Największą chyba nadzieję na znalezienie dowodu dają nowe lub całkowicie odmienne sposoby myślenia o tym problemie. Jak się już wielokrotnie przekonaliśmy, do przełomów w pracach nad wielkimi problemami dochodzi często wtedy, gdy komuś udaje się je powiązać z jakimiś zupełnie innymi obszarami matematyki. Dobrym tego przykładem jest wielkie twierdzenie Fermata: gdy udało się je wyrazić w postaci pytania dotyczącego krzywych eliptycznych, prace ruszyły ostro z miejsca.

Taktyka de Brangesa wydaje się obecnie wątpliwa, ale ze strategicznego punktu widzenia jego podejście jest jak najbardziej właściwe. Opiera się ono na pewnej uwadze Davida Hilberta z 1912 roku. Niezależnie od Hilberta to samo zauważył George Pólya. Fizyk Edmund Landau poprosił Pólyę, by wyjaśnił mu, jakie fizyczne podstawy przemawiają za prawdziwością hipotezy Riemanna. W 1982 roku Pólya wspominał, że znalazł rozwiązanie tak postawionego problemu: miejsca zerowe funkcji dzeta powinny być związane z wartościami własnymi tak zwanego operatora samo-sprzężonego. Są to pewne wartości charakterystyczne związane ze specjalnymi rodzajami przekształceń. W fizyce kwantowej, która jest jednym z ważnych zastosowań tej teorii, liczby te określają poziomy energetyczne rozważanych układów i z pewnego standardowego i łatwego twierdzenia wynika, że wartości własne tego specjalnego operatora są zawsze rzeczywiste. Jak

powiedzieliśmy, hipotezę Riemanna można wyrazić równoważnie jako stwierdzenie, że wszystkie miejsca zerowe funkcji ksi są rzeczywiste. Gdyby jakiś operator samosprzężony miał wartości własne takie same jak miejsca zerowe funkcji ksi, to można by prosto pokazać, że wynika z tego prawdziwość hipotezy Riemanna. Pólya nie opublikował tych przemyśleń – nie udało mu się zapisać takiego operatora i dopóki ktoś tego nie dokona, dopóty dyskusja na ten temat jest obiecywaniem gruszek na wierzbie. Jednak w 1950 roku Selberg udowodnił swój „wzór śladowy", który łączy geometrię powierzchni z wartościami własnymi związanego z nią operatora. Fakt ten sprawił, że pomysł Pólyi nabrał nieco bardziej realnych kształtów.

W 1972 roku Hugh Montgomery odwiedził Instytut Studiów Zaawansowanych w Princeton. W swoich badaniach zwrócił uwagę na pewne zaskakujące własności statystyczne nietrywialnych miejsc zerowych funkcji dzeta i wspomniał o tym fizykowi Freemanowi Dysonowi, który z miejsca dostrzegł ich podobieństwo do cech statystycznych losowych macierzy hermitowskich – są one innym specjalnym rodzajem operatora i wykorzystuje się je do opisu układów kwantowych, takich jak jądra atomowe. W 1999 roku Alain Connes wymyślił wzór śladowy podobny do wzoru Selberga, z którego prawdziwości powinna wynikać prawdziwość uogólnionej hipotezy Riemanna. W tym samym roku fizycy Michael Berry i Jon Keating zasugerowali, że poszukiwany operator można by uzyskać przez skwantowanie pewnego dobrze znanego z fizyki klasycznej pojęcia związanego z pędem. Tak powstała hipoteza Berry'ego, którą można uważać za szczególną wersję hipotezy Hilberta–Pólyi.

Wszystkie te koncepcje wiążące hipotezę Riemanna z podstawowymi działami fizyki matematycznej są wspaniałe. Pokazują, że jakiś postęp w pracach nad tą hipotezą może się dokonać w pozornie niezwiązanych z nią gałęziach matematyki. Możemy więc wciąż mieć nadzieję, że kiedyś uda się w końcu rozstrzygnąć, czy hipoteza ta jest prawdziwa. Wysiłki te nie doprowadziły jednak jeszcze do żadnego przełomu, który uprawniałby nas do wyrażenia opinii, że rozwiązanie jest już

w zasięgu ręki. Hipoteza Riemanna pozostaje jedną z najbardziej zdumiewających i irytujących zagadek matematycznych.

Niedawno doszedł jeszcze jeden powód, dla którego warto udowodnić hipotezę Riemanna: wysoka nagroda pieniężna.

Nagrody Nobla nie przyznaje się w dziedzinie matematyki. Najbardziej prestiżową nagrodą matematyczną jest Medal Fieldsa, a mówiąc ściśle, Międzynarodowy Medal za Wybitne Odkrycia Matematyczne. Wyróżnienie to zostało tak nazwane na cześć kanadyjskiego matematyka Johna Fieldsa, który ustanowił odznaczenie w swoim testamencie. Co cztery lata na Międzynarodowym Kongresie Matematyków maksymalnie czterech wyróżniających się w świecie młodych (to znaczy takich, którzy nie ukończyli jeszcze 40 lat) badaczy otrzymuje złoty medal i nagrodę pieniężną, która obecnie wynosi 15 tysięcy dolarów. Wśród matematyków Medal Fieldsa ma taką samą rangę jak Nagroda Nobla.

Wielu matematyków uważa, że to dobrze, iż w ich dziedzinie nie przyznaje się Nagrody Nobla. Obecnie laureat Nagrody Nobla otrzymuje nieco ponad milion dolarów i taka kwota bez trudu mogłaby wpłynąć na wybór celów badawczych, a nawet doprowadzić do sporów o pierwszeństwo odkryć. Jednak brak wysokiej nagrody w dziedzinie matematyki może wpływać negatywnie na opinię społeczeństwa na temat wartości i przydatności tej dziedziny nauki. Nietrudno dojść do wniosku, że skoro nikt nie chce płacić za takie badania, to nie mogą być one zbyt wiele warte.

Niedawno ustanowiono dwie nowe nagrody matematyczne o dużym znaczeniu. Pierwszą z nich jest Nagroda Abela przyznawana corocznie przez Norweską Akademię Nauk i Literatury, nosząca imię wielkiego norweskiego matematyka Nielsa Henrika Abela. Drugim nowym wyróżnieniem jest siedem nagród milenijnych ustanowionych przez Instytut Matematyczny Claya. Instytut Claya został założony przez Landona Claya i jego żonę Lavinię. Landon Clay jest amerykańskim biznesmenem działającym na rynku funduszy inwestycyjnych, który szczególnie upodobał sobie matematykę.

W 1999 roku założył w Cambridge w stanie Massachusetts nową fundację matematyczną, która organizuje spotkania, przyznaje fundusze na badania, przygotowuje wykłady popularnonaukowe i zarządza przyznawaną corocznie nagrodą za osiągnięcia badawcze.

W 2000 roku sir Michael Atiyah i John Tate, czołowi matematycy Wielkiej Brytanii i Stanów Zjednoczonych, ogłosili, że Instytut Matematyczny Claya ustanowił nową nagrodę, która ma zachęcić uczonych do poszukiwania rozwiązań siedmiu najważniejszych nierozstrzygniętych zagadnień matematycznych. Zagadnienia te stały się powszechnie znane jako problemy milenijne. Rozwiązanie każdego z nich, opublikowane w uznanym czasopiśmie i zweryfikowane przez recenzentów, zostanie nagrodzone czekiem na milion dolarów. Wszystkie te problemy, wybrane starannie przez czołowych matematyków świata, zwracają uwagę społeczności uczonych na niektóre z najważniejszych nierozwiązanych zagadnień. Fakt, że za ich rozwiązanie zostanie przyznana wysoka nagroda pieniężna, jest bardzo jasnym przesłaniem dla całego społeczeństwa: matematyka jest cenna. Wszyscy zajmujący się tymi problemami doskonale zdają sobie sprawę, że ich rozwiązanie będzie miało olbrzymie znaczenie naukowe, którego nie sposób wycenić, ale nagroda pieniężna na pewno jest dobrą dodatkową zachętą, by się tymi zagadnieniami zająć. Najbardziej znanym problemem milenijnym, który z całej siódemki najdłużej czeka na rozwiązanie, jest udowodnienie hipotezy Riemanna. Jest to jedyne zagadnienie, które figuruje na liście Hilberta z 1900 roku i na liście problemów milenijnych. Pozostałe sześć problemów milenijnych omówimy w rozdziałach od 10 do 15. Matematycy nie mają obsesji na punkcie nagród i na pewno zajmowaliby się hipotezą Riemanna również bez perspektywy otrzymania pokaźnej kwoty. Nowa, obiecująca idea byłaby wystarczającą motywacją do pracy.

Nie wolno jednak zapominać, że hipotezy, nawet te znane od dawna, nie muszą być prawdziwe. Obecnie większość matematyków skłania się ku opinii, że hipotezę Riemanna uda się w końcu udowodnić. Nieliczni jednak uważają, że hipoteza ta może się ostatecznie okazać nieprawdziwa, że gdzieś w gąszczu bardzo wielkich liczb czai się miejsce zerowe, które nie znajduje się na

prostej krytycznej. Jeśli taki kontrprzykład istnieje, to jest on najprawdopodobniej naprawdę liczbowo duży.

Jednak w wypadku najnowszych badań matematycznych opinie poszczególnych osób nie mają zbyt dużego znaczenia. To bardzo dobrze, gdy ekspert ma jakieś przeczucia, ale w przeszłości wielokrotnie już się zdarzało, że prowadziły nas one na manowce. Powszechnie akceptowane poglądy wcale nie muszą być prawdziwe. Littlewood, jeden z wielkich specjalistów od analizy zespolonej, w 1962 roku stwierdził kategorycznie, że ma pewność, iż hipoteza Riemanna jest nieprawdziwa, dodając, że trudno sobie wyobrazić jakiś powód, dla którego miałoby być inaczej. Kto ma rację? Pożyjemy, zobaczymy.

64 Spotyka się również odmianę tej całki, w której wyrażenie $1/\ln t$ całkuje się od 2 do *x*, zamiast od 0 do *x*. W ten sposób unikamy problemów rachunkowych w punkcie $t = 0$, w którym wartość $\ln t$ nie jest zdefiniowana. Czasami zapis $\mathrm{Li}(x)$ oznacza właśnie ten wariant, a funkcję zdefiniowaną w tekście oznacza się wówczas jako $\mathrm{li}(x)$.

65 Imię Pafnutij jest rzadko spotykane. Zainspirowało to Philipa Davisa do napisania dziwnej, ale wciągającej książki: *The Thread: a Mathematical Yarn* (*Wątek: matematyczna opowieść*), Harvester Press, 1983.

66 Wynika to z interesującego wzoru Riemanna

$$\zeta(1-s) = 2^{1-s}\pi^{-s}\sin\left(\frac{\pi(1-s)}{2}\right)\Gamma(s)\zeta(s)$$

gdzie Γ(s) jest klasyczną funkcją o nazwie gamma, określoną dla wszystkich liczb zespolonych *s*. Prawa strona równości jest określona dla wszystkich liczb zespolonych *s*, których część rzeczywista jest większa od 1.

67 Bernhard Riemann, *Über die Anzahl der Primzahlen unter einer gegebenen Grösse* (*O liczbie liczb pierwszych mniejszych od danej wielkości*), „Monatsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin", listopad 1859.

68 Riemann zdefiniował również podobną funkcję

$$\Pi(x)=\pi(x)+\frac{1}{2}\pi\left(x^{1/2}\right)+\frac{1}{3}\pi\left(x^{1/3}\right)+\frac{1}{4}\pi\left(x^{1/4}\right)+\cdots$$

która zlicza potęgi liczb pierwszych, a nie same liczby pierwsze. Na jej podstawie możemy wyznaczyć π(*x*). Następnie Riemann wyprowadził dokładny wzór na taką zmodyfikowaną funkcję pozwalający wyrazić ją za pomocą logarytmu całkowego i drugiej podobnej całki:

$$\Pi(x)=\mathrm{Li}(x)-\sum_{\rho}\mathrm{Li}\left(x^{\rho}\right)+\int_{x}^{\infty}\frac{dt}{t\left(t^{2}-1\right)\ln t}$$

Symbol Σ oznacza tu sumę po wszystkich liczbach ρ, dla których $\zeta(\rho)$ jest równe zero, z wyjątkiem ujemnych liczb parzystych.

69 Na przykład $x+\sqrt{x}$ jest asymptotycznie równe *x*. Iloraz tych wyrażeń wynosi:

$$\left(x+\sqrt{x}\right)\Big/x=1+1\Big/\sqrt{x}$$

Gdy *x* rośnie, zwiększa się również wartość $\sqrt{x}$, a zatem $1/\sqrt{x}$ zmierza do 0 i nasz iloraz zmierza do 1. Różnica jednak cały czas wynosi $\sqrt{x}$ i staje się coraz większa, w miarę jak rośnie wartość *x*. Gdy na przykład *x* osiąga wartość jednego biliona, to $\sqrt{x}$ wynosi milion.

70 Stała Eulera jest granicą wyrażenia

$$1+\frac{1}{2}+\frac{1}{3}+\cdots\frac{1}{n}-\ln n$$

gdy *n* zmierza do nieskończoności.

71 Douglas A. Stoll, Patrick Demichel, *The impact of* $\zeta(s)$ *complex zeros on* π(*x*) *for* $x < 10^{10^{13}}$ (*Wpływ zespolonych miejsc zerowych* $\zeta(s)$ *na* π(*x*) *dla* $x < 10^{10^{13}}$), „Mathematics of Computation" 2011, tom 276, s. 2381–2394.

72 Zob.: http://empslocal.ex.ac.uk/people/staff/mrwatkin/zeta/RHproofs.htm.

73 J. Brian Conrey, Xian-Jin Li, *A note on some positivity conditions related to zeta- and L-functions* (*Uwaga na temat pewnych warunków dodatniości związanych z funkcjami dzeta i L*), http://arxiv.org/abs/math.NT/9812166.

10. Jaki kształt ma sfera?

Hipoteza Poincarégo

Henri Poincaré był jednym z największych matematyków końca XIX wieku. Był do pewnego stopnia ekscentrykiem, ale także zręcznym kombinatorem. Został członkiem francuskiej instytucji naukowej pod nazwą Bureau des Longitudes, która zajmowała się usprawnieniem metod nawigacji, pomiaru czasu oraz mierzenia Ziemi i innych planet. W ramach prac w tej organizacji przedstawił propozycję układu międzynarodowych stref czasowych. Prace te skłoniły go również do zastanowienia się nad fizyką czasu i doszedł do wniosków wyprzedzających niektóre odkrycia szczególnej teorii względności Einsteina. Poincaré odcisnął swoje piętno we wszystkich dziedzinach matematyki, od teorii liczb po fizykę matematyczną.

W szczególności był jednym z twórców topologii, czyli matematyki przekształceń ciągłych. Właśnie podczas tych pionierskich badań w 1904 roku natrafił na pewne proste z pozoru pytanie i poniewczasie zrozumiał, że we wcześniejszej pracy niejawnie przyjął pewne związane z nim założenie, choć nie potrafił znaleźć odpowiedniego dowodu. „To pytanie wymagałoby zbyt dalekiego odejścia od głównego tematu” – napisał, choć w istocie owo stwierdzenie nie oddaje całej wagi problemu: prawda jest taka, że zastanawiając się nad nim, nie zdołał dojść *dokądkolwiek*. Chociaż wyraził ten problem w formie pytania, stał się on powszechnie znany jako hipoteza Poincarégo, ponieważ wszyscy spodziewali się, że odpowiedź na nie brzmi: „tak”. Jest to kolejny z siedmiu problemów milenijnych ogłoszonych przez Instytut Claya – i słusznie, ponieważ okazało się, że jest to jeden z najbardziej zagadkowych problemów

w całej topologii. Na pytanie Poincarégo odpowiedzi udzielił dopiero w końcu w 2002 roku młody Rosjanin Grigorij Perelman. Znalezienie rozwiązania wymagało wykorzystania olbrzymiej ilości nowych idei i metod, do tego stopnia, że społeczność matematyków potrzebowała kilku lat, by przegryźć się przez przedstawiony dowód i potwierdzić jego poprawność.

Za te osiągnięcia Perelmanowi przyznano Medal Fieldsa, najbardziej prestiżowe wyróżnienie matematyczne, ale uczony odmówił jego przyjęcia. Nie chciał rozgłosu. Za udowodnienie hipotezy Poincarégo Instytut Claya wyznaczył nagrodę w wysokości miliona dolarów, ale Perelman ją odrzucił. Nie chciał też pieniędzy. Pragnął jedynie, żeby jego prace zaakceptowała społeczność matematyków. W końcu tak się stało, choć niestety, ze zrozumiałych powodów, zajęło to trochę czasu. Poza tym trudno się spodziewać akceptacji, której nie towarzyszyłyby rozgłos i nagrody. Jednak te nieuniknione konsekwencje sukcesu stały w sprzeczności z dość skrytą naturą Perelmana.

Mówiliśmy już o topologii przy okazji omawiania twierdzenia o czterech barwach i wtedy użyłem wyświechtanego frazesu: „gumowa geometria". Geometria euklidesowa zajmuje się prostymi, okręgami, długościami i kątami. Wszystko dzieje się w niej na płaszczyźnie lub, w bardziej zaawansowanej wersji, w trójwymiarowej przestrzeni. Płaszczyzna przypomina nieskończoną kartkę papieru i ma z nią jedną podstawową cechę wspólną: nie rozciąga się, nie kurczy ani nie wygina. Można wprawdzie zwinąć kartkę w rulon, może się też ona nieco skurczyć lub rozciągnąć, zwłaszcza gdy wylejemy na nią kawę, ale nie zdołamy owinąć nią kuli tak, by nie powstały żadne zagięcia. Z matematycznego punktu widzenia płaszczyzna euklidesowa jest sztywna. W geometrii Euklidesa dwa obiekty – trójkąty, kwadraty, koła – są takie same wtedy, gdy można przekształcić jeden w drugi za pomocą przekształceń izometrycznych, czyli takich, które nie zmieniają odległości.

Co by się stało, gdybyśmy zamiast papieru użyli jakiejś elastycznej powierzchni? Takiej, która się rozciąga i wygina, a nawet, jeśli się postaramy, daje się ścisnąć. Na takiej giętkiej powierzchni długości i kąty nie mają trwałego znaczenia. A jeśli powierzchnia ta będzie *wystarczająco* sprężysta, to swoje trwałe znaczenie utracą nawet trójkąty, kwadraty i koła. Na gumowej powierzchni można zniekształcić trójkąt, dodając mu jeszcze jeden róg. Można go nawet zmienić w koło (zob. ryc. 36). Zatem, bez względu na to, jakie są pojęcia takiej gumowej geometrii, możemy być pewni, że nie obejmują one tradycyjnych, euklidesowych definicji.

Ryc. 36. Topologiczne przekształcenie trójkąta w koło

Mogłoby się wydawać, że geometria na gumowej powierzchni powinna być tak zmienna, iż nie będzie w niej nic stałego, a w takiej sytuacji trudno się spodziewać, że dałoby się w jej ramach cokolwiek udowodnić. A jednak tak nie jest. Narysujmy trójkąt i umieśćmy w jego środku punkt. Jeśli teraz będziemy rozciągać i odkształcać powierzchnię tak, że trójkąt stanie się kołem, to jedna cecha naszego rysunku nie ulegnie zmianie: punkt pozostanie wewnątrz figury. To prawda, że teraz znajduje się wewnątrz koła, nie trójkąta, ale nie jest *na zewnątrz*. Aby go przesunąć poza figurę, musielibyśmy rozedrzeć naszą powierzchnię, a to byłoby złamanie zasad gry.

Jest jeszcze jedna cecha, która nie ulega zmianie w obliczu zniekształceń. Trójkąt jest krzywą zwyczajną zamkniętą. Jest nieprzecinającą się linią, której końce są ze sobą połączone. Cyfra osiem jest krzywą zamkniętą, ale nie zwyczajną, ponieważ przecina samą siebie. Gdy zaczniemy zniekształcać gumową powierzchnię, trójkąt może zmieniać kształt, ale zawsze pozostanie krzywą

zwyczajną zamkniętą. Bez rozdzierania powierzchni w żaden sposób nie zdołamy zmienić go na przykład w figurę w kształcie ósemki.

W topologii trójwymiarowej cała przestrzeń staje się elastyczna. Nie przypomina jednak kawałka gumy, który powraca do pierwotnego kształtu, gdy go puścimy. Bardziej jest podobna do żelu, który pozwala się odkształcać bez żadnego oporu. Przestrzeń topologiczna jest nieskończenie odkształcalna. Możemy wybrać sobie obszar o rozmiarze ziarnka ryżu i powiększyć go do wielkości Słońca. Możemy ciągnąć za wybrane miejsca danego obszaru, rozciągając przestrzeń w długie macki, tak że zacznie przypominać ośmiornicę. Jedyne, czego nam nie wolno, to przerywać ciągłości. Nie wolno nam rozdzierać przestrzeni ani dokonywać jakichkolwiek odkształceń prowadzących do oderwania od siebie pobliskich punktów.

Jakie cechy figur w przestrzeni zdołają przetrwać wszelkie ciągłe odkształcenia? Na pewno nie są to długość, pole powierzchni czy objętość. Jednak zasupłanie już tak. Jeśli zawiążemy na krzywej węzeł i połączymy jej końce, tworząc pętlę, to taki supeł nigdy nie zdoła z niej uciec. Jakkolwiek byśmy odkształcali przestrzeń, nie zdołamy usunąć węzła z naszej krzywej. Mamy więc nowy rodzaj geometrii, w której ważnymi pojęciami są dość mgliste określenia: „wnętrze", „zamknięta", „zwyczajna" i „zawiązana". Ta nowa geometria ma budzącą szacunek nazwę: topologia. Może się wydawać dość nieprzystępna, a nawet absurdalna, ale okazuje się, że jest ona jednym z głównych działów badań matematyki XX wieku i pozostaje równie istotna w XXI stuleciu. A jedną z najważniejszych osób, którym zawdzięczamy powstanie tej dziedziny, jest Poincaré.

Historia topologii rozpoczęła się niemal sto lat przed pojawieniem się Poincarégo, w 1813 roku. Szwajcarski matematyk Simon Antoine Jean Lhuillier nie wzbudził za życia zbyt dużego zainteresowania swą osobą w świecie matematyki, mimo że odrzucił pokaźną kwotę, jaką jeden z krewnych obiecał mu, jeśli wstąpi w służbę Kościoła. Lhuillier wybrał jednak matematykę. Zajął się zupełnie niemodną gałęzią, a mianowicie twierdzeniem Eulera o wielościanach. W rozdziale 4 spotkaliśmy się już z tym dziwnym, pozornie niezwiązanym z niczym

wzorem: jeśli wielościan ma S ścian, W wierzchołków i K krawędzi, to $S - K + W = 2$. Lhuillier poświęcił znaczną część swego życia na analizę różnych odmian tego wzoru i patrząc z perspektywy czasu, możemy dostrzec, że zrobił kluczowy krok w kierunku topologii, gdy odkrył, że w niektórych wypadkach wzór Eulera jest błędny. O tym, czy wzór ten jest poprawny, decyduje rodzaj kształtu danego wielościanu.

Wzór Eulera działa poprawnie dla wielościanów pozbawionych otworów, czyli takich, które można narysować na powierzchni kuli lub zniekształcić je w sposób ciągły, tak że staną się sferą. Jednak w wypadku wielościanów z otworami wzór ten zawodzi. Rama obrazu wykonana z drewnianych listew o prostokątnym przekroju ma na przykład 16 ścian, 32 krawędzie i 16 wierzchołków, zatem $S - K + W = 0$. Lhuillier zmodyfikował wzór Eulera tak, by obejmował również takie bardziej egzotyczne wielościany: jeśli wielościan ma g otworów, to wtedy $S - K + W = 2 - 2g$. Był to pierwszy ważny niezmiennik topologiczny, jaki udało się odkryć. Niezmiennikiem topologicznym nazywamy wielkość związaną z przestrzenią, która nie ulega zmianie, gdy sama przestrzeń podlega odkształceniom ciągłym. Niezmiennik Lhuilliera pozwala w ścisły sposób policzyć, ile otworów ma dana powierzchnia, bez konieczności definiowania, czym jest „otwór". To bardzo dobrze się składa, ponieważ opisanie otworu jest dość skomplikowane. Otwór nie jest częścią powierzchni, nie jest też obszarem leżącym poza nią. Wydaje się, że jest wyrazem tego, jak dana powierzchnia jest osadzona w otaczającej ją przestrzeni. Jednak odkrycie Lhuilliera pokazuje, że *liczbę* otworów można uważać za nieodłączną cechę powierzchni, niezależną od otaczającej ją przestrzeni. Nie musimy definiować otworów, by móc je potem policzyć – mówiąc szczerze, lepiej tego nie robić.

Drugą po Lhuillierze kluczową postacią prehistorii topologii jest Gauss, który podczas prac w różnych innych podstawowych dziedzinach matematyki odkrył kilka innych niezmienników topologicznych. Jego prace z zakresu analizy zespolonej, a w szczególności dowód, że każde równanie wielomianowe ma przynajmniej jedno rozwiązanie zespolone, sprawiły, iż zaczął się zastanawiać nad indeksem punktu względem krzywej na

płaszczyźnie, czyli nad tym, ile razy owija się ona wokół danego punktu. Z kolei problemy z dziedziny elektryczności i magnetyzmu kazały mu analizować liczbę opleceń dwóch krzywych zamkniętych, czyli ile razy jedna z krzywych oplata się wokół drugiej. Te i inne przykłady sprawiły, że Gauss zaczął rozważać, czy mogłaby istnieć jakaś nieodkryta dotąd gałąź matematyki, która pozwoliłaby w systematyczny sposób zrozumieć cechy jakościowe figur geometrycznych. Nie napisał na ten temat żadnego artykułu, ale wspominał o tym w listach i rękopisach.

Tymi przemyśleniami podzielił się ze swoim studentem Johannem Listingiem i asystentem Augustusem Möbiusem. Wspomnieliśmy już o wstędze Möbiusa – powierzchni mającej tylko jedną stronę i jedną krawędź, którą opisał w artykule z 1865 roku (zob. ryc. 9). Möbius zwrócił uwagę na fakt, że chociaż określenie „mająca tylko jedną stronę" jest intuicyjnie jasne, to trudno je opisać w sposób ścisły, i zaproponował zastosowanie pokrewnej własności, którą można opisać z zachowaniem pełnej ścisłości. Własnością tą jest orientowalność. Powierzchnia jest orientowalna wtedy, gdy można ją pokryć siecią trójkątów ze strzałkami krążącymi wokół ich boków w taki sposób, że na wszystkich bokach współdzielonych przez dwa trójkąty strzałki w każdym z nich wskazują w przeciwnych kierunkach. Właśnie tak się stanie, jeśli narysujemy taką sieć na płaszczyźnie i przyjmiemy, że w każdym trójkącie strzałki biegną na przykład zgodnie z kierunkiem ruchu wskazówek zegara. Na wstędze Möbiusa nie można skonstruować takiej sieci.

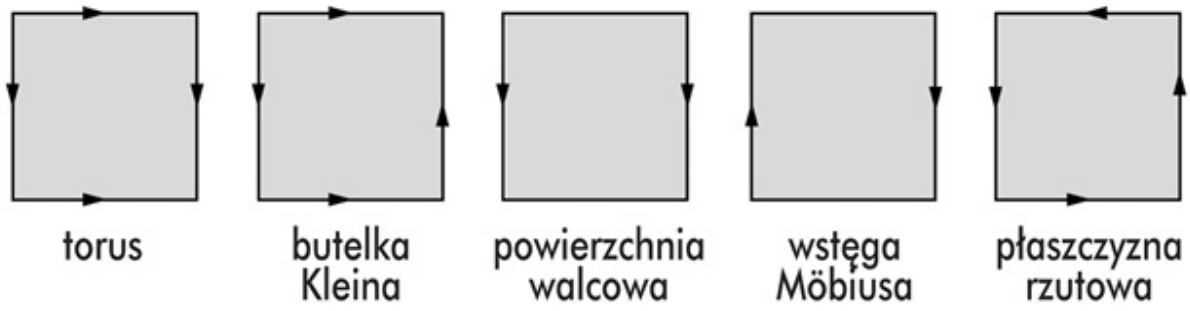
Listing opublikował swoją pierwszą pracę poświęconą topologii już wcześniej, w 1847 roku, w książce zatytułowanej *Vorstudien zur Topologie* (*Wykłady na temat topologii*) i był to pierwszy tekst naukowy, w którym użyto tego słowa. Listing posługiwał się tą nazwą nieoficjalnie już od mniej więcej dziesięciu lat. W tamtym okresie używano jeszcze innej nazwy na określenie nowej dziedziny badań, a mianowicie łacińskiego zwrotu *analysis situs* – *analiza położenia* – ale ostatecznie wyszedł on z użycia. W dziele Listinga nie znajdziemy zbyt wielu ważnych odkryć, ale opisał w nim jedno z podstawowych podejść – metodę pokrywania powierzchni siecią trójkątów. Już w 1861 roku, na cztery lata przed Möbiusem, opisał

wstęgę Möbiusa i badał jej spójność, czyli to, czy można daną przestrzeń podzielić na dwie lub więcej rozdzielnych części. Prace Listinga stały się punktem wyjścia dla badań wielu matematyków, między innymi Walthera von Dycka, którzy określili pełną topologiczną klasyfikację powierzchni przy założeniu, że są one zamknięte (nie mają krawędzi) i zwarte (mają skończone rozmiary). Okazało się, że każda orientowalna powierzchnia jest topologicznie równoważna sferze z doczepioną do niej skończoną liczbą g uchwytów. Liczbę g nazywamy genusem powierzchni i to właśnie jej użył Lhuillier w swoim niezmienniku. Jeśli $g = 0$, to mamy sferę, natomiast dla $g > 0$ otrzymujemy torus mający g otworów. Podobny ciąg powierzchni, zaczynający się od najprostszej powierzchni nieorientowalnej, czyli płaszczyzny rzutowej, stanowi klasyfikację wszystkich powierzchni nieorientowalnych. Metodę tę rozszerzono również na wszystkie powierzchnie z brzegiem. Brzeg jest zamkniętą pętlą i jedyną dodatkową informacją, jaką musimy mieć w tej sytuacji, jest liczba tych pętli.

Hipoteza Poincarégo nabierze większego sensu, jeśli najpierw przyjrzymy się metodom stosowanym w klasyfikacji powierzchni. Opisując wcześniej topologię, mówiliśmy o odkształcaniu figur wykonanych z gumy lub żelu i podkreśliliśmy, że możemy stosować jedynie przekształcenia *ciągłe*. Jak na ironię, jedna z głównych technik stosowanych w topologii polega na czynności, która wydaje się przekształceniem nieciągłym – na pocięciu danego obiektu na części. Ciągłość można jednak odtworzyć dzięki liście reguł opisujących, które kawałki łączą się ze sobą i w jaki sposób. Z przykładem takiego podejścia spotkaliśmy się już w rozdziale 4, gdy zdefiniowaliśmy torus, utożsamiając ze sobą przeciwległe boki kwadratu (zob. ryc. 12).

Utożsamiając ze sobą punkty, które wydają się różne, możemy przedstawiać skomplikowane przestrzenie topologiczne za pomocą prostszych składników. Kwadrat jaki jest, każdy widzi, ale kwadrat z regułami utożsamiania może być torusem, butelką Kleina, powierzchnią walcową, wstęgą Möbiusa lub płaszczyzną rzutową –

zależnie od wyboru reguł (zob. ryc. 37). Gdy więc wyjaśnialiśmy przekształcenia ciągłe za pomocą rozciągania i zakrzywiania gumowej powierzchni, wymagaliśmy więcej, niż to konieczne. Okazuje się, że na etapie pośrednim możemy również przeciąć badaną powierzchnię, pod warunkiem jednak iż na koniec połączymy ze sobą krawędzie dokładnie tak, jak były, lub określimy reguły oznaczające praktycznie to samo. Jeśli chodzi o topologów, to samo *określenie* reguł sklejenia ze sobą krawędzi jest dla nich równie dobre jak wprowadzenie tych reguł w życie. Nie wolno tylko zapomnieć o istnieniu reguł we wszystkich działaniach, jakie się później przeprowadza.



Ryc. 37. Pięć różnych przestrzeni topologicznych, które można uzyskać dzięki utożsamieniu ze sobą przeciwległych brzegów kwadratu na różne sposoby

Pierwszym krokiem do przeprowadzenia klasyfikacji danej powierzchni metodą klasyczną jest narysowanie na niej sieci trójkątów. Następnie należy ją przeciąć wzdłuż tylu krawędzi, by dało się te trójkąty rozłożyć płasko i uzyskać wielokąt. Trzeba przy tym określić reguły sklejania wynikające z tego, jak pocięliśmy powierzchnię, a także ustalić, w jaki sposób należy utożsamić ze sobą poszczególne boki wielokąta, co jest równoznaczne z odtworzeniem początkowej powierzchni. W tym momencie wszystkie cechy powierzchni interesujące z punktu widzenia topologii zawarte są w sposób niejawny w regułach sklejania. Klasyfikacji dokonuje się, przeprowadzając operacje algebraiczne na tych

regułach i przekształcając je do postaci odpowiadającej torusowi o *g* otworach lub jednej z analogicznych powierzchni nieorientowalnych. Współczesna topologia dysponuje również innymi metodami prowadzącymi do otrzymania takiego samego wyniku, ale i tak często stosuje się konstrukcje typu „wytnij i sklej". Taką metodę można bez większych trudności uogólnić na przestrzenie o dowolnej liczbie wymiarów, jest ona jednak zbyt ograniczona, by bez dodatkowej pomocy pozwalała na uzyskanie klasyfikacji przestrzeni topologicznych o wyższej liczbie wymiarów.

Około roku 1900 Poincaré przekształcił wcześniejsze metody poświęcone topologii powierzchni w znacznie bardziej ogólną technikę, którą można stosować w przestrzeniach o dowolnej liczbie wymiarów. Głównym motywem jego badań były poszukiwania niezmienników topologicznych: liczb lub wzorów algebraicznych związanych z przestrzeniami, które nie zmieniają się, gdy sama przestrzeń podlega odkształceniom ciągłym. Jeśli dwie przestrzenie mają różne niezmienniki, to jednej nie da się przekształcić w drugą, a zatem są topologicznie różne.

Punktem wyjścia jego prac było uogólnienie topologicznego niezmiennika Lhuilliera $S - K + W$ na przestrzenie o większej liczbie wymiarów, opracowane w 1870 roku przez włoskiego matematyka Enrica Bettiego, które obecnie, jakże niesprawiedliwie, nazywa się charakterystyką Eulera. Betti zauważył, że największa liczba krzywych zamkniętych, jakie można narysować na powierzchni o genusie *g* bez dzielenia jej na rozłączne fragmenty, wynosi $g - 1$. Jest to jeszcze inny sposób opisu topologicznego danej powierzchni. Uogólnił swój pomysł, wprowadzając „liczby spójności" dowolnej liczby wymiarów, które Poincaré nazwał liczbami Bettiego i tak już zostało. *K*-ta liczba Bettiego jest liczbą *k*-wymiarowych otworów w danej przestrzeni.

Poincaré przekształcił liczby spójności Bettiego w bardziej wrażliwy niezmiennik, nazywany homologią, obdarzony znacznie bogatszą strukturą algebraiczną. O samej homologii powiemy więcej w rozdziale 15. W tej chwili wystarczy, jeśli będziemy wiedzieli, że

polega ona na przyjrzeniu się wielowymiarowym „ścianom” pewnego rodzaju sieci i sprawdzeniu, które z nich stanowią brzeg topologicznego dysku. Dysk, w przeciwieństwie do torusa, nie ma dziur, możemy mieć więc pewność, że w każdym zbiorze ścian tworzących taki brzeg nie ma żadnych otworów. Możemy też wykryć otwory, porównując zbiory ścian nietworzących brzegu ze zbiorami ścian brzegowych. W ten sposób uzyskamy ciągi niezmienników przestrzeni, które nazywamy jej grupami homologii. Słowa „grupa” użyto tutaj w znaczeniu wywodzącym się z algebry abstrakcyjnej – oznacza ono mianowicie, że jeśli połączymy ze sobą dwa dowolne obiekty danej grupy w sposób zgodny z kilkoma zgrabnymi regułami algebraicznymi, to uzyskamy jakiś inny obiekt należący do tej samej grupy. Trochę więcej na ten temat powiemy później, gdy to pojęcie będzie nam bardziej potrzebne. Dla każdego wymiaru od 0 do *n* istnieje jedna taka grupa i dla każdej przestrzeni otrzymujemy ciąg topologicznych niezmienników o przeróżnych fascynujących własnościach algebraicznych.

Listing dokonał klasyfikacji wszystkich powierzchni topologicznych – czyli dwuwymiarowych przestrzeni. Oczywistym następnym krokiem było przejście do trzech wymiarów, a najprostszym obiektem, od którego można zacząć to dzieło, wydawała się sfera, czyli powierzchnia kuli. Należy pamiętać, że wnętrze sfery nie jest jej częścią – jest jedynie wynikiem tego, jak zazwyczaj osadzamy powierzchnię sferyczną w przestrzeni. Mamy zatem jedynie powierzchnię równoważną topologicznie powierzchni piłki. Możemy sobie wyobrażać sferę jako pustą w środku piłkę o nieskończenie cienkiej ściance.

„Właściwy” trójwymiarowy odpowiednik sfery, nazywany trójwymiarową hipersferą, *nie jest* wypełnioną w środku piłką. Wypełniona piłka jest wprawdzie trójwymiarowa, ale ma brzeg, swoją powierzchnię, która jest sferą. Sfera natomiast nie ma brzegu, a zatem nie może go mieć również jej trójwymiarowy odpowiednik. Trójwymiarową hipersferę najprościej można zdefiniować, naśladując własności geometryczne zwyczajnej sfery. W ten sposób otrzymujemy przestrzeń, którą dość trudno jest sobie wyobrazić. Nie mogę tu pokazać jej trójwymiarowego modelu, ponieważ

trójwymiarowa hipersfera – choć faktycznie ma tylko trzy wymiary – nie daje się osadzić w zwyczajnej trójwymiarowej przestrzeni. Wymaga użycia przestrzeni o czterech wymiarach.

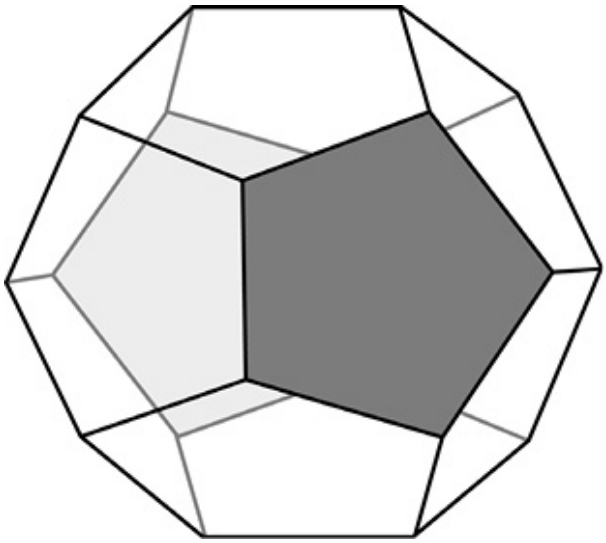
Zwyczajną sferę jednostkową w trójwymiarowej przestrzeni tworzą wszystkie punkty leżące w odległości 1 jednostki od ustalonego punktu – jej środka. Podobnie trójwymiarową jednostkową hipersferę w czterowymiarowej przestrzeni tworzą wszystkie punkty znajdujące się w odległości 1 jednostki od jej środka. W układzie współrzędnych możemy zapisać jej wzór, definiując odległość w oparciu o uogólnione twierdzenie Pitagorasa[74]. Mówiąc ogólniej, trójwymiarowa hipersfera w *dowolnej* przestrzeni jest topologicznie równoważna jednostkowej trójwymiarowej hipersferze, tak jak wszelkie powyginane odmiany jednostkowej dwuwymiarowej sfery są topologicznie dwuwymiarowymi sferami – i oczywiście to samo obowiązuje w wyższych wymiarach.

Być może takie wyjaśnienie nie usatysfakcjonuje wszystkich, którzy spodziewali się jakiegoś bardziej geometrycznego opisu, spróbujmy więc zdefiniować hipersferę jeszcze inaczej. Trójwymiarową hipersferę można przedstawić jako kulę, której całą powierzchnię utożsamiono z pojedynczym punktem. Jest to kolejny przykład reguły sklejania – i tym razem mamy do czynienia z przypadkiem podobnym do sposobu pozwalającego przekształcić zwyczajne płaskie koło w dwuwymiarową sferę. Jeśli wokół obwodu koła wyciętego z materiału rozciągniemy sznurek, a następnie mocno go ściągniemy, tak jak gdybyśmy zaciskali worek, to uzyskamy obiekt topologicznie równoważny dwuwymiarowej sferze. Teraz przeprowadźmy podobną operację na kuli, ale jak poprzednio, nie próbujmy sobie wyobrażać wyniku – po prostu pomyślmy o kuli i zastosujmy w myślach odpowiednie reguły sklejania.

Tak czy inaczej, Poincaré zainteresował się trójwymiarową hipersferą, ponieważ jest to chyba najprostsza trójwymiarowa przestrzeń topologiczna o skończonym rozmiarze, która nie ma brzegu. W 1900 roku opublikował artykuł, w którym stwierdził, że grupy homologii są tak silnym niezmiennikiem, iż umożliwiają topologiczne scharakteryzowanie trójwymiarowej hipersfery. Mówiąc

ściśle, zauważył, że każda trójwymiarowa przestrzeń topologiczna mająca takie same grupy homologii jak trójwymiarowa hipersfera jest topologicznie równoważna trójwymiarowej hipersferze (czyli innymi słowy, można ją w sposób ciągły przekształcić w trójwymiarową hipersferę). Jednak do 1904 roku odkrył, że twierdzenie to jest nieprawdziwe. Istnieje przynajmniej jeszcze jedna trójwymiarowa przestrzeń niebędąca hipersferą, mająca jednak takie same grupy homologii jak trójwymiarowa hipersfera. Przestrzeń ta jest świadectwem triumfu podejścia wykorzystującego reguły sklejania i dowodem na to, że to wcale nie trójwymiarowa hipersfera jest potrzebna do skonstruowania nowego niezmiennika, który siłą rzeczy powinien być znacznie silniejszy od homologii.

Zacznijmy od przestrzeni. Nazywamy ją dwunastościenną przestrzenią Poincarégo, ponieważ we współczesnych konstrukcjach wykorzystuje się właśnie dwunastościan foremny. Poincaré nie wiedział o istnieniu tego związku z dwunastościanem – uzyskał tę przestrzeń, sklejając ze sobą dwa torusy w bardzo zawiły sposób. Konstrukcję z wykorzystaniem dwunastościanu opisali po raz pierwszy w 1933 roku, 21 lat po śmierci Poincarégo, Herbert Seifert i Constantin Weber – to podejście jest znacznie łatwiejsze do zrozumienia. Jest ono podobne do konstruowania torusa przez sklejenie ze sobą przeciwległych boków kwadratu. Jak pamiętamy, nie próbowaliśmy wtedy *wykonywać* takich sklejeń. Wystarczyło, że przyjęliśmy, iż odpowiadające sobie punkty brzegu są jednym i tym samym punktem. Teraz postąpmy tak samo z przeciwległymi ścianami dwunastościanu (zob. ryc. 38).



Ryc. 38. Aby uzyskać dwunastościenną przestrzeń Poincarégo, należy skleić ze sobą wszystkie przeciwległe pary ścian dwunastościanu foremnego (takie jak dwie ściany zaznaczone szarym kolorem), przekręcając je przy tym tak, by do siebie pasowały.

Pitagorejczycy badali dwunastościany już 2500 lat temu. Brzeg dwunastościanu foremnego tworzy 12 pięciokątów foremnych połączonych ze sobą w taki sposób, by powstała mniej więcej sferyczna klatka. W każdym rogu takiej bryły spotykają się trzy pięciokąty. Mając taką bryłę, sklejamy ze sobą przeciwległe ściany… trzeba jeszcze tylko pamiętać o przekręcie. Dosłownym. Przed sklejeniem każdą ścianę należy obrócić o odpowiedni kąt. Jest to najmniejszy kąt pozwalający na dopasowanie odpowiednich ścian – kąt ten wynosi 36 stopni. Możemy sobie wyobrazić, że jest to nieco bardziej złożona wersja przepisu na wstęgę Möbiusa, który wymagał przekręcenia brzegu o 180 stopni przed sklejeniem obu końców paska.

Tak właśnie wygląda owa przestrzeń. Przyjrzymy się teraz niezmiennikowi. Wszystko, o czym tu mówimy, nie jest bujaniem w obłokach – przyda nam się to do zrozumienia hipotezy Poincarégo.

Poincaré nazwał swój nowy niezmiennik grupą podstawową. Obecnie matematycy wciąż posługują się tą nazwą, ale często nazywają go również (pierwszą) grupą homotopii. Homotopia to konstrukcja geometryczna dostarczająca informacji na temat topologii danej przestrzeni, którą można przeprowadzić całkowicie w jej wnętrzu. Dokonuje się tego z wykorzystaniem abstrakcyjnej struktury algebraicznej, którą nazywamy grupą. Grupa jest zbiorem obiektów matematycznych cechujących się tym, że połączenie dowolnych dwóch z nich daje inny obiekt należący do tej grupy. To prawo łączenia – często nazywane mnożeniem lub dodawaniem, nawet gdy nie chodzi o zwyczajne operacje arytmetyczne – musi mieć kilka prostych i naturalnych własności. Jeśli nazwiemy to działanie dodawaniem, to główne wymagania będą następujące:

- Grupa musi zawierać element zachowujący się jak zero: jeśli dodamy go do dowolnego innego elementu, uzyskamy ten sam element.
- Każdy element grupy musi mieć swój ujemny odpowiednik – jeśli je do siebie dodamy, to otrzymamy zero.
- Gdy dodajemy do siebie trzy elementy grupy, to kolejność wykonywania tych działań nie może mieć znaczenia. Innymi słowy, $(a + b) + c = a + (b + c)$. Własność tę nazywamy prawem łączności.

Jedyną własnością algebraiczną, której się w tym wypadku *nie wymaga* (chociaż czasami jest ona również spełniona), jest przemienność[75]: $a + b = b + a$.

Grupa podstawowa Poincarégo jest swego rodzaju uproszczonym szkieletem przestrzeni. Jest niezmiennikiem topologicznym – przestrzenie równoważne topologicznie mają tę samą grupę podstawową. Aby spróbować to zrozumieć i być może odtworzyć motywację Poincarégo, zobaczymy, jak to działa na przykładzie okręgu, wykorzystując porównanie użyte kiedyś przez Gaussa. Wyobraźmy sobie mrówkę żyjącą na okręgu, dla której jest on całym wszechświatem. W jaki sposób mogłaby ona ustalić, jaki kształt ma

jej wszechświat? Czy byłaby w stanie odróżnić okrąg, powiedzmy, od prostej? Pamiętajmy cały czas, że nasza mrówka nie może wyjść poza swój wszechświat, spojrzeć na niego z oddali i przekonać się, że jest okrągły. Może jedynie przemieszczać się w jego wnętrzu, jakiekolwiek by ono było. W szczególności mrówka nie może stwierdzić, że wszechświat jest zakrzywiony, ponieważ w jej świecie odpowiedniki promieni świetlnych również są uwięzione na okręgu. Nie zagłębiajmy się tu jednak w różne kwestie praktyczne, takie jak problem przechodzenia obiektów przez siebie – nasze porównanie jest dość luźne.

Mrówka może odkryć, jaki kształt ma jej wszechświat, na kilka różnych sposobów. My skupimy się tutaj na metodzie, którą można uogólnić na dowolną przestrzeń topologiczną. Przyjmijmy na potrzeby naszych rozważań, że mrówka jest punktem. Mieszka przy przystanku autobusowym, który również jest punktem. Każdego dnia wychodzi z domu, wsiada do autobusu (który oczywiście również jest punktem) i dociera z powrotem do domu. Najbardziej oczywista jest trasa linii autobusowej numer 0, która polega na tym, że autobus po prostu stoi na przystanku i nigdzie się nie rusza. Gdy mrówka ma ochotę na bardziej interesującą przejażdżkę, wsiada do autobusu linii 1, który okrąża jednokrotnie wszechświat w kierunku przeciwnym do ruchu wskazówek zegara i wraca na ten sam przystanek. Autobus linii 2 wykonuje dwa okrążenia, trasa linii 3 polega na wykonaniu trzech okrążeń i tak dalej: każdej liczbie dodatniej odpowiada jedna linia autobusowa krążąca w kierunku przeciwnym do ruchu wskazówek zegara. Są również autobusy oznakowane liczbami ujemnymi, które jeżdżą w przeciwnym kierunku. Autobus linii –1 robi jedno okrążenie w kierunku zgodnym z ruchem wskazówek zegara, autobus linii –2 robi dwa takie okrążenia i tak dalej.

Mrówka szybko zauważa, że dwie przejażdżki z rzędu autobusem linii 1 są w zasadzie tym samym co jedna podróż dwójką, a trzy przejażdżki jedynką to w zasadzie to samo co jedna wyprawa autobusem linii 3. Podobnie przejażdżka autobusem z numerem 5 i przesiadka do autobusu linii 8 jest w zasadzie tym samym co jedna podróż autobusem numer 13. W istocie, dla dowolnych dwóch liczb dodatnich przejażdżka autobusem o numerze linii równym pierwszej

z tych liczb i zaraz potem przesiadka do autobusu linii odpowiadającej drugiej liczbie jest równoważna jednej podróży autobusem, którego numer linii jest równy sumie obu liczb.

Następny krok jest nieco bardziej złożony. Taki sam związek jest *niemal* prawdziwy dla autobusów linii ujemnych i linii zerowej. Podróż autobusem linii 0 i przesiadka do autobusu linii 1 jest bardzo podobna do przejażdżki autobusem linii 1. Istnieje jednak pewna subtelna różnica. W czasie podróży 0 + 1 autobus linii 0 stoi przez pewien określony czas na przystanku, co nie zdarza się w trakcie pojedynczej przejażdżki jedynką. Wprowadźmy więc pojęcie o groźnie brzmiącej nazwie: homotopia (które po grecku oznacza „to samo miejsce"). Dwie pętle są ze sobą homotopijne, jeśli jedną z nich można przekształcić w drugą w sposób ciągły. Jeśli pozwolimy, żeby homotopie wpływały na plan podróży autobusowej, to sprawimy, że czas, jaki mrówka spędza, stojąc na przystanku w autobusie linii 0, będzie się stopniowo kurczył, aż całkowicie zniknie. W takiej sytuacji znika różnica między podróżą 0 + 1 a przejazdem autobusem 1, zatem „z dokładnością do homotopii" wynikiem jest jeden przejazd autobusem linii 1. Innymi słowy, równanie linii autobusowych 0 + 1 = 1 jest prawdziwe, jednak nie dla samych przejazdów, ale dla klas ich homotopii.

A co w przypadku przejazdu autobusem linii 1 i następnie przesiadki do autobusu linii –1? Chcielibyśmy, żeby taka wyprawa była równoważna przejażdżce autobusem linii 0, ale tak nie jest. W czasie takiej podróży mrówka wykonuje najpierw jedno okrążenie w kierunku przeciwnym do ruchu wskazówek zegara, a następnie robi jeszcze jedno kółko, tym razem w drugą stronę. Nie ma wątpliwości, że to co innego niż siedzenie przez cały ten czas w autobusie linii 0 stojącym na przystanku. Zatem 1 + (–1), czyli 1 – 1, nie jest równe 0. Z pomocą przychodzi nam jednak ponownie homotopia: połączenie przejazdów autobusami 1 i –1 jest homotopijne z przejażdżką autobusem 0. Aby zrozumieć dlaczego, przypuśćmy, że mrówka przemierza połączoną trasę autobusów linii 1 i –1 własnym samochodem, ale po pierwszym okrążeniu, tuż przed dojechaniem do przystanku, zawraca i jedzie do domu w przeciwnym kierunku. Taka trasa jest bardzo podobna do połączonej przejażdżki

dwoma autobusami, pomija jedynie maleńki odcinek podróży. Zatem pierwotny przejazd dwoma autobusami „skurczył się” w sposób ciągły do nieco krótszej podróży samochodem. Mrówka może skrócić tę drogę jeszcze bardziej, zawracając nieco wcześniej. Jeśli będzie tak dalej postępowała, zawracając za każdym razem odrobinę wcześniej, dojdzie w końcu do sytuacji, w której będzie siedziała w nieruchomym samochodzie zaparkowanym na przystanku. Taka operacja kurczenia również jest homotopią i pokazuje, że przejazd autobusem linii 1 z przesiadką do autobusu linii –1 jest homotopijny z przejazdem autobusem linii 0. Innymi słowy, 1 + (–1) = 0 dla klas homotopii przejazdów.

Teraz każdy, kto zna algebrę, może bez trudu udowodnić, że przejazd dowolnym autobusem i następnie przesiadka do innego autobusu jest homotopijny z przejazdem autobusem linii będącej sumą obydwu numerów autobusów. Jest to prawdą dla autobusów dodatnich, ujemnych i autobusu zerowego. Jeśli więc dodamy do siebie przejazdy autobusowe – a raczej klasy homotopii przejazdów autobusowych – to otrzymamy grupę. W istocie jest to pewna dobrze znana grupa. Jej elementami są liczby całkowite (numery linii autobusowych), a jej działaniem – dodawanie. Przyjęło się ją oznaczać symbolem $\mathbb{Z}$, od niemieckiego słowa *Zahl* (liczba).

Trzeba się dużo bardziej napracować, żeby udowodnić, że w okrągłym wszechświecie *dowolna* podróż samochodem – nawet taka, w której długie odcinki pokonuje się na wstecznym biegu, często się zawraca i jeździ tam i z powrotem po tym samym odcinku drogi – jest homotopijna z jedną ze standardowych tras linii autobusowych. Co więcej, trasy linii autobusowych o różnych numerach nie są ze sobą homotopijne. Udowodnienie tego wymaga nieco głębszej znajomości teorii, ale podstawową ideą jest tu gaussowski indeks punktu względem krzywej. Indeks ten zlicza całkowitą liczbę okrążeń, jakie dana trasa wykonuje po okręgu w kierunku przeciwnym do ruchu wskazówek zegara[76]. Na jego podstawie można ustalić, z którą trasą linii autobusowej jest homotopijna dana trasa przejazdu samochodem.

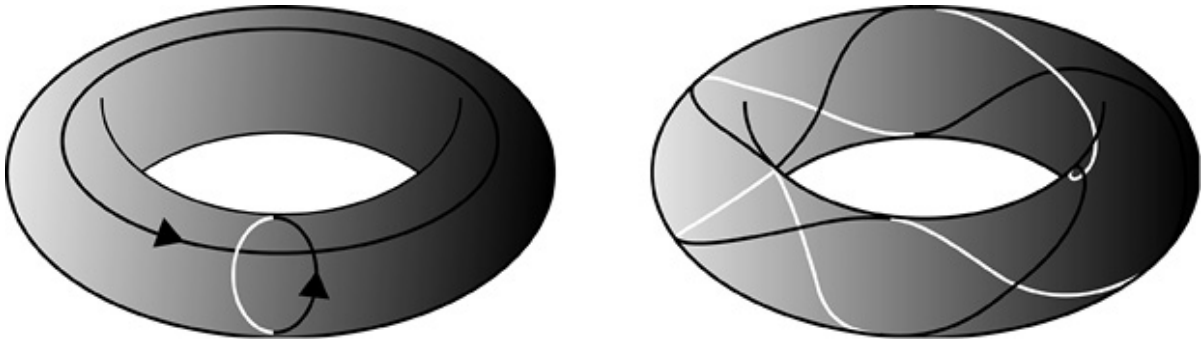
Po uzupełnieniu wszystkich szczegółów taki opis dowodzi, że grupą podstawową okręgu jest ta sama grupa $\mathbb{Z}$, którą tworzy zbiór liczb całkowitych z dodawaniem. Aby dodać przejazdy, wystarczy dodać ich indeksy punktu względem krzywej. Mrówka mogłaby wykorzystać ten niezmiennik topologiczny do odróżnienia wszechświata kołowego od, powiedzmy, linii prostej. Na prostej każda trasa, bez względu na to, jak zawiła, musi w pewnym momencie osiągnąć maksymalną odległość od punktu wyjścia. Możemy teraz skracać tę trasę w sposób ciągły, kurcząc jednakowo wszystkie odległości mierzone od punktu wyjścia – najpierw do 99 procent, potem do 98 i tak dalej. Zatem na prostej *każda* trasa jest homotopijna z zerem – z sytuacją, gdy się nigdzie nie ruszamy. Grupa podstawowa prostej ma tylko jeden element: 0. Jej własności algebraiczne są trywialne: 0 + 0 = 0. Nic więc dziwnego, że nazywa się ją grupą trywialną, i ponieważ różni się ona od grupy wszystkich liczb całkowitych, mrówka może stwierdzić, czy żyje na prostej, czy na okręgu.

Jak powiedzieliśmy, są jeszcze inne metody, ale stosując takie podejście, mrówka może znaleźć odpowiedź na dręczące ją pytanie dzięki wykorzystaniu pojęcia grupy podstawowej Poincarégo.

Idźmy dalej. Przypuśćmy teraz, że nasza mrówka żyje na pewnej powierzchni. Tak jak poprzednio, jest ona całym jej światem – mrówka nie może wyjść poza tę powierzchnię, żeby sprawdzić, jakiego jest ona rodzaju. Czy może ustalić, jaka jest topologia jej wszechświata? A w szczególności czy jest w stanie zauważyć różnicę między sferą i torusem? Odpowiedź na oba pytania jest twierdząca i by znaleźć te rozwiązania, należy się posłużyć tą samą metodą jak w wypadku wszechświata kołowego – trzeba wsiąść do autobusu i udać się podróż, która zaczyna się i kończy tuż obok domu. Aby dodać do siebie dwa przejazdy, należy je wykonać jeden po drugim. Trasa zerowa to „stanie na przystanku”, a trasa odwrotna do danej to ta sama trasa, tylko przejechana w przeciwnym kierunku. Przy takich założeniach uzyskujemy grupę, pod warunkiem że działamy na klasach homotopii przejazdów. Jest to grupa podstawowa danej powierzchni. W porównaniu z wszechświatem kołowym mamy teraz

większą swobodę w planowaniu tras przejazdu i przekształcania ich w sposób ciągły w inne trasy, ale również w takiej sytuacji nasza podstawowa idea wciąż działa.

Grupa podstawowa w dalszym ciągu jest niezmiennikiem topologicznym i mrówka może ją wykorzystać do ustalenia, czy żyje na sferze, czy na torusie. Jeśli jej wszechświat jest sferyczny, to bez względu na to, jaką trasą pojedzie, zawsze będzie można ją stopniowo odkształcić do trasy zerowej – czyli postoju na przystanku. Nie jest to prawdą w odniesieniu do wszechświata w kształcie torusa. Niektóre trasy da się w nim odkształcić do zera, ale nie tę zawierającą okrążenie przez otwór w środku (zob. ryc. 39, po lewej). Takie stwierdzenie wymaga dowodu, ale jego dostarczenie nie jest trudne. Linie autobusowe na torusie również mają standardową numerację, teraz jednak numer każdego autobusu tworzą pary liczb całkowitych (*m*, *n*). Pierwsza liczba *m* mówi o tym, ile dana trasa robi okrążeń przez otwór w środku torusa, natomiast druga liczba, *n*, określa, ile razy autobusy tej linii okrążają cały torus. Na rycinie 39, po prawej stronie, pokazano trasę (5, 2), która przebiega pięć razy przez otwór środkowy, okrążając jednocześnie torus dwukrotnie. Aby dodać do siebie dwie trasy, należy zsumować odpowiadające sobie liczby, na przykład: (3, 6) + (2, 4) = (5, 10). Grupą podstawową torusa jest grupa $\mathbb{Z}^2$ *par* liczb całkowitych.



Ryc. 39. Po lewej: Trasa linii autobusowych (1, 0) i (0, 1) na torusie. Po prawej: Trasa linii autobusowej (5, 2). Linie narysowane kolorem szarym biegną z tyłu torusa.

Każda przestrzeń topologiczna ma grupę podstawową zdefiniowaną w taki sam sposób z wykorzystaniem tras – fachowo nazywanych pętlami – które zaczynają się i kończą w tym samym punkcie. Poincaré wymyślił grupę podstawową, by udowodnić, że jego przestrzeń dwunastościenna nie jest trójwymiarową hipersferą, mimo iż obie przestrzenie mają takie same niezmienniki homologiczne. Już jego oryginalna konstrukcja świetnie nadaje się do obliczenia takiej grupy podstawowej, ale współczesna metoda polegająca na „przekręceniu i sklejeniu" będzie jeszcze lepsza. Okazuje się, że rozwiązaniem jest grupa 120 elementów związanych z dwunastościanem. Dla porównania – grupa podstawowa trójwymiarowej hipersfery ma tylko jeden element – pętlę zerową. Zatem przestrzeń dwunastościenna *nie jest* topologicznie równoważna sferze, mimo że obie mają tę samą homologię – tym samym Poincaré dowiódł, że jego stwierdzenie z 1900 roku jest błędne.

Następnie zaczął się zastanawiać nad swoim nowym niezmiennikiem: czy jest on brakującym składnikiem charakterystyki topologicznej trójwymiarowej hipersfery? Może każda trójwymiarowa przestrzeń z taką samą grupą podstawową jak trójwymiarowa hipersfera – czyli z grupą będącą grupą trywialną – musi w istocie *być* trójwymiarową hipersferą? Sformułował to w postaci pytania sugerującego negatywną odpowiedź: „Rozważmy zwartą trójwymiarową rozmaitość [czyli przestrzeń topologiczną] *V* pozbawioną brzegu. Czy to możliwe, by grupa podstawowa *V* była grupą trywialną, jeśli *V* nie jest [topologicznie] równoważna trójwymiarowej hipersferze?". Pozostawił to pytanie bez odpowiedzi, ale powszechne przekonanie, że odpowiedź jest oczywista – przy takim sformułowaniu aż się prosi, by odpowiedzieć „nie" – stało się wkrótce znane jako hipoteza Poincarégo. Równie szybko stała się ona jednym z najbardziej uciążliwych otwartych pytań w topologii.

Określenie „trywialna grupa podstawowa" jest równoważne stwierdzeniu, że „każdą pętlę można w sposób ciągły przekształcić w punkt". Nie tylko trójwymiarowa hipersfera ma taką własność. To

samo dotyczy *n*-wymiarowej hipersfery. Możemy więc wysunąć taką samą hipotezę dla hipersfery o dowolnej liczbie wymiarów. Takie sformułowanie jest nazywane *n*-wymiarową hipotezą Poincarégo. Z twierdzenia klasyfikującego powierzchnie wynika, że jest ona prawdziwa dla przypadku $n = 2$. Przez ponad pięćdziesiąt lat nikomu nie udało się zrobić kolejnego kroku.

W 1961 roku Stephen Smale wykorzystał pewną sztuczkę używaną przy klasyfikacji powierzchni i zastosował ją w wyższych wymiarach. Jednym ze sposobów na uzyskanie torusa o *g* otworach jest dodanie do sfery *g* uchwytów – wypustek przypominających ucho filiżanki lub kubka. Smale uogólnił tę konstrukcję na dowolną liczbę wymiarów. Zastanawiał się, jak można zmieniać uchwyty bez zmiany topologii przestrzeni, i doszedł do wniosku potwierdzającego hipotezę Poincarégo dla wszystkich wymiarów większych lub równych 7. Jego dowód nie działa dla mniejszej liczby wymiarów, ale inni matematycy wymyślili, jak można go poprawić: John Stallings przedstawił taki dowód dla sześciu wymiarów, a Christopher Zeeman dla pięciu. Jednak pewien kluczowy krok, znany jako sztuczka Whitneya, zawodzi dla trzech i czterech wymiarów, ponieważ w tych przestrzeniach nie ma wystarczająco dużo miejsca na przeprowadzenie wymaganych operacji i nikomu nie udało się znaleźć jakiejś działającej alternatywy tego podejścia. Pojawiło się ogólne przeświadczenie, że w przestrzeniach trój- i czterowymiarowych topologia może być odmienna niż w innych przestrzeniach.

W 1982 roku Michael Freedman zadał kłam temu powszechnemu przekonaniu, udowadniając hipotezę Poincarégo dla czterech wymiarów bez korzystania ze sztuczki Whitneya. Jego dowód jest niezwykle skomplikowany, ale poprawny. Zatem po 50 latach zastoju i 20 gorączkowych badań topologom udało się w końcu udowodnić hipotezę Poincarégo dla wszystkich wymiarów z wyjątkiem jednego – akurat tego, który Poincaré zawarł w swoim pytaniu. Odniesione sukcesy są imponujące, ale metody użyte do przeprowadzenia tych dowodów niewiele nam mówią o przypadku przestrzeni trójwymiarowej. Potrzebne było jakieś inne podejście.

Wyjście z tego impasu umożliwił dopiero zestaw przypominający tradycyjną listę elementów stroju panny młodej: coś starego, coś nowego, coś pożyczonego i – naciągając nieco porównanie – coś niebieskiego. Czymś starym było ponowne odwiedzenie znanego od dawna obszaru topologii, który po gorączce prac poświęconych przestrzeniom o większej liczbie wymiarów uznano powszechnie za wyeksploatowany – chodzi o topologię powierzchni. Czymś nowym było ponowne przemyślenie kwestii klasyfikacji powierzchni z punktu widzenia, który wydawał się początkowo zupełnie dziwny – a mianowicie z punktu widzenia geometrii klasycznej. Czymś pożyczonym był przepływ Ricciego, który powstał w wyniku prac nad opisem matematycznym ogólnej teorii względności Einsteina. A czymś niebieskim – rozmyślania „o niebieskich migdałach" – snucie daleko idących hipotez opartych na odrobinie intuicji i solidnej dawce nadziei.

Jak pamiętamy, możemy uporządkować listę wszystkich powierzchni orientowalnych bez brzegu, ponieważ każda z nich jest topologicznie równoważna torusowi z pewną liczbą otworów. Liczbę tę nazywamy genusem danej powierzchni i gdy wynosi ona zero, to mamy do czynienia ze sferą bez uchwytów – czyli po prostu ze sferą. Przypomina nam to, że wśród wszystkich sfer topologicznych istnieje jedna powierzchnia, którą możemy nazwać archetypem wszystkich sfer. Chodzi mianowicie o sferę jednostkową w przestrzeni euklidesowej. Zapomnijmy na chwilę o całym tym rozciąganiu gumowych obiektów. Zaraz do niego wrócimy, ale skupmy się na razie na starej dobrej sferze euklidesowej. Ma ona wiele dodatkowych własności matematycznych, które wynikają ze sztywności geometrii euklidesowej. Najważniejsza z nich jest krzywizna, którą można opisać ilościowo. Każdemu punktowi powierzchni geometrycznej można przypisać liczbę, która będzie miarą zakrzywienia powierzchni w jego otoczeniu. W przestrzeni euklidesowej sfera jest jedyną powierzchnią zamkniętą, której krzywizna jest taka sama w każdym punkcie – i do tego jest dodatnia.

To dziwne, ponieważ stała krzywizna nie jest własnością topologiczną. Co jeszcze dziwniejsze, sfera nie jest jedynym tego

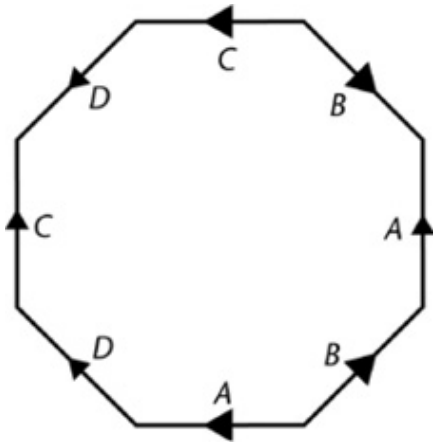
typu przypadkiem. Istnieje również jedna standardowa powierzchnia geometryczna, którą możemy nazwać archetypem torusa. Aby ją uzyskać, należy narysować kwadrat na płaszczyźnie i utożsamić ze sobą jego przeciwległe boki (zob. ryc. 12). Jeśli narysujemy to w trójwymiarowej przestrzeni i zwiniemy kwadrat tak, by odpowiednie boki zetknęły się ze sobą, otrzymamy powierzchnię, która wygląda na zakrzywioną. Jednak z punktu widzenia samej powierzchni możemy wykonywać wszystkie działania na kwadracie, pamiętając tylko o regułach sklejenia. Kwadrat ma naturalną strukturę geometryczną – jest mianowicie obszarem płaszczyzny euklidesowej. Płaszczyzna również ma stałą krzywiznę, tyle tylko że jest ona wszędzie *zerowa*. Torus o takiej geometrii też ma zerową krzywiznę, dlatego nazywamy go *płaskim torusem*. Brzmi to niemal jak oksymoron, ale mrówka żyjąca na płaskim torusie, uzbrojona w linijkę i kątomierz, po przeprowadzeniu odpowiednich pomiarów może zaświadczyć, że lokalna geometria jest na nim taka sama jak na płaszczyźnie.

Osiemnastowieczni specjaliści od geometrii usiłowali zrozumieć aksjomat Euklidesa o prostych równoległych i postanowili wyprowadzić go z pozostałych założeń podstawowych. Ich próby raz za razem kończyły się porażką i w końcu uświadomili sobie, że takie wyprowadzenie jest niemożliwe. Istnieją trzy różne rodzaje geometrii i w każdym z nich spełnione są wszystkie warunki sformułowane przez Euklidesa, z wyjątkiem postulatu równoległości. Geometrie te noszą nazwy: euklidesowa (geometria płaszczyzny, w której postulat równoległości jest spełniony), eliptyczna (geometria powierzchni sfery z pewnymi dodatkami, w której każde dwie proste przecinają się, a zatem proste równoległe nie istnieją) oraz hiperboliczna (w której niektóre proste się nie przecinają i dlatego równoległość nie jest cechą jednoznaczną). Klasyczni matematycy interpretowali te geometrie jako geometrie zakrzywionych przestrzeni. Geometria euklidesowa odpowiada krzywiźnie zerowej, eliptyczna (sferyczna) – stałej krzywiźnie dodatniej, natomiast hiperboliczna – stałej krzywiźnie ujemnej.

Powiedzieliśmy przed chwilą, jak można uzyskać pierwsze dwie z wymienionych geometrii – są to geometrie na sferze i płaskim

torusie. Z punktu widzenia klasyfikacji powierzchni są to torusy z *g* otworami dla *g* = 0 oraz 1. Brakuje nam jeszcze geometrii hiperbolicznej. Czy każdy torus z dowolną liczbą otworów ma naturalną strukturę geometryczną bazującą na określonym wielokącie w przestrzeni hiperbolicznej, w którym utożsamiono ze sobą niektóre jego boki? To zaskakujące, ale odpowiedź brzmi: „tak" dla *dowolnej* wartości *g* większej lub równej 2. Na rycinie 40 pokazano przykład dla *g* = 2 bazujący na ośmiokącie. Nie będziemy się tu zagłębiać w kwestie geometrii hiperbolicznej i utożsamienia tej powierzchni z torusem z dwoma otworami, ale można to wszystko pokazać. Różne wartości genusu powstają z różnych wielokątów i można w ten sposób uzyskać każdą wartość *g*. Używając matematycznego żargonu, powiedzielibyśmy, że każdy torus z dwoma lub więcej otworami ma naturalną strukturę hiperboliczną. Możemy więc teraz sformułować na nowo listę standardowych powierzchni:

- Sfera, *g* = 0: geometria eliptyczna.
- Torus, *g* = 1: geometria euklidesowa.
- Torus z *g* otworami, *g* = 2, 3, 4, …: geometria hiperboliczna.



Ryc. 40. Konstrukcja torusa z dwoma otworami za pomocą ośmiokąta, w którym utożsamiono ze sobą pary jego boków (AA, BB, CC, DD)

Mogłoby się wydawać, że wylaliśmy tu dziecko z kąpielą, ponieważ topologia ma być przecież związana z geometrią gumową, a nie sztywną. Teraz jednak możemy bez trudu wstawić z powrotem całą elastyczność geometrii gumowej. Tradycyjna, sztywna geometria służy tu jedynie do *zdefiniowania* standardowych powierzchni. Pozwala uzyskać proste opisy, które przypadkiem mają dodatkową sztywną strukturę. Teraz wystarczy, że usuniemy tę sztywność i pozwolimy, żeby przestrzeń *stała się* elastyczna jak guma. Niech się dowolnie odkształca, nawet tak, jak nie mogła, gdy miała sztywną strukturę. Uzyskamy w ten sposób powierzchnie równoważne powierzchniom standardowym z punktu widzenia topologii, ale nie z punktu widzenia izometrii. Z twierdzenia o klasyfikacji powierzchni wynika, że w ten sposób możemy otrzymać każdą powierzchnię topologiczną.

Topolodzy mieli świadomość istnienia tego związku między geometrią i twierdzeniem o klasyfikacji powierzchni, ale wyglądało to raczej na zabawny zbieg okoliczności, który bez wątpienia musi mieć związek z dość ograniczonymi możliwościami, jakie mamy w dwóch wymiarach. Wszyscy wiedzieli, że przypadek trójwymiarowy jest znacznie bogatszy i w szczególności przestrzenie o stałej krzywiźnie nie wyczerpują wszystkich możliwości. Dopiero jednak William Thurston, jeden z największych specjalistów od geometrii na świecie, uświadomił sobie, że sztywna geometria może w dalszym ciągu być przydatna w trójwymiarowej topologii. Już wcześniej pojawiło się kilka wskazówek: choćby to, że trójwymiarowa hipersfera Poincarégo ma naturalną eliptyczną (sferyczną) geometrię, która wynika z jej definicji. Chociaż standardowy dwunastościan istnieje w przestrzeni euklidesowej, to kąt między sąsiednimi ścianami wynosi mniej niż 120 stopni, a zatem trzy takie kąty nie zakreślają pełnego okręgu. Aby temu zaradzić, musimy nadąć dwunastościan tak, by jego ściany stały się lekko wypukłe – tym sposobem jego naturalna geometria staje się sferyczna, nie euklidesowa. Podobnie trójkąty na sferze również są wypukłe. Trójwymiarowy torus powstały w wyniku utożsamienia ze sobą przeciwległych ścian sześcianu ma płaską –

czyli euklidesową – geometrię, podobnie jak jego dwuwymiarowy odpowiednik. Max Dehn i inni matematycy odkryli kilka trójwymiarowych przestrzeni topologicznych o naturalnej geometrii hiperbolicznej.

Thurston zaczął dostrzegać przejawy istnienia jakiejś ogólnej teorii, ale by jej znalezienie było w ogóle możliwe, trzeba było wprowadzić jeszcze dwie zmiany. Po pierwsze, należało rozszerzyć zakres trójwymiarowych geometrii. Thurston spisał sensowne warunki i udowodnił, że spełnia je dokładnie osiem geometrii. Trzy z nich to znane już geometrie klasyczne: sferyczna, euklidesowa i hiperboliczna. Dwie kolejne przypominają walec – są płaskie w jednym kierunku, a zakrzywione w pozostałych dwóch. Zakrzywione fragmenty mogą mieć krzywiznę dodatnią (dwuwymiarowa sfera) albo ujemną (płaszczyzna hiperboliczna). W końcu istnieją jeszcze trzy dodatkowe, dość zawiłe geometrie.

Po drugie, niektóre trójwymiarowe przestrzenie nie pasowały do żadnej z tych ośmiu geometrii. Rozwiązaniem okazało się pocięcie ich na kawałki. Jeden fragment mógł mieć strukturę geometrii sferycznej, inny hiperbolicznej i tak dalej. Aby to miało sens, cięcie należało przeprowadzić w ściśle określony sposób, tak żeby ponowne poskładanie kawałków dostarczyło użytecznych informacji. Na szczęście okazało się, że w wielu wypadkach taka operacja jest możliwa. W 1982 roku, dając świadectwo ogromnej wyobraźni geometrycznej, Thurston wysunął hipotezę geometryzacyjną: *każdą* trójwymiarową przestrzeń można pociąć w sposób w zasadzie jednoznaczny na fragmenty mające naturalną strukturę geometryczną odpowiadającą jednej z ośmiu możliwych geometrii. Dowiódł również, że jeśli jego hipoteza geometryzacyjna jest prawdziwa, to hipoteza Poincarégo będzie prostą, wynikającą z niej konsekwencją.

Tymczasem pojawiła się możliwość zaatakowania hipotezy Poincarégo z jeszcze innej strony. Ta nowa linia ataku również miała charakter geometryczny, także opierała się na krzywiznach, ale wywodziła się z zupełnie innej dziedziny – z fizyki matematycznej.

Gauss, Riemann i specjaliści z włoskiej szkoły geometrii rozwinęli ogólną teorię przestrzeni zakrzywionych (nazywanych rozmaitościami) ze zdefiniowanym pojęciem odległości, co niesamowicie rozszerzyło euklidesową i klasyczną nieeuklidesową geometrię. Krzywizna nie musiała już być stała – mogła zmieniać się gładko między różnymi punktami. Na przykład bryła w kształcie psiej kości ma dodatnią krzywiznę na obu końcach, ale między nimi krzywizna jest ujemna i przy przejściu między sąsiednimi obszarami zmienia się łagodnie. Krzywiznę opisuje się ilościowo, wykorzystując do tego narzędzia matematyczne zwane tensorami. Około roku 1915 Albert Einstein uświadomił sobie, że to właśnie tensory krzywizny są tym, czego potrzebował, by rozszerzyć szczególną teorię względności, która mówiła tylko o przestrzeni i czasie, i zaproponować ogólną teorię względności, obejmującą również grawitację. W tej teorii pole grawitacyjne przedstawia się jako krzywiznę przestrzeni i równania pola Einsteina opisują, jak związana z nią miara krzywizny, czyli tensor krzywizny, zmienia się pod wpływem rozkładu materii. Można powiedzieć, że krzywizna przestrzeni *przepływa* z biegiem czasu – Wszechświat lub jakaś jego część spontanicznie zmienia swój kształt.

Richard Hamilton, specjalista od geometrii riemannowskiej, zauważył, że tę samą sztuczkę można zastosować w bardziej ogólnym przypadku i takie podejście może prowadzić do uzyskania dowodu hipotezy Poincarégo. Pomysł polega na użyciu jednej z najprostszych miar krzywizny, nazywanej krzywizną Ricciego od nazwiska włoskiego matematyka Gregoria Ricciego-Curbastra. Hamilton przedstawił równanie opisujące, jak krzywizna Ricciego powinna się zmieniać w czasie – jest to tak zwany przepływ Ricciego (zwany też potokiem Ricciego). Skonstruował to równanie w taki sposób, by krzywizna stopniowo rozlewała się jak najbardziej równomiernie. Przypomina to trochę nasze porównanie z kotem pod dywanem z rozdziału 4, z tą różnicą, że chociaż kot nadal nie może wyjść spod dywanu, to teraz jednak potrafi rozłożyć się płasko na całej powierzchni, tworząc równą warstwę. (Jak widać, w tej sytuacji można się posłużyć wyłącznie kotem topologicznym).

W dwuwymiarowym przypadku możemy na przykład wyjść od powierzchni w kształcie gruszki (zob. ryc. 41). Na jednym z jej końców znajduje się obszar o dużej dodatniej krzywiźnie, po przeciwnej stronie również mamy krzywiznę dodatnią, ale już nie tak silną, a między nimi – pas o krzywiźnie ujemnej. Działanie przepływu Ricciego polega w zasadzie na przenoszeniu krzywizny z silnie zakrzywionego końca (i w mniejszym zakresie z drugiego, grubszego końca) na zakrzywiony ujemnie pas tak długo, aż cała ujemna krzywizna zostanie pochłonięta. Na tym etapie powstaje wypukła powierzchnia o dodatniej krzywiźnie w każdym miejscu. Przepływ Ricciego nie przestaje jednak działać i nadal przemieszcza krzywiznę – zabiera jej trochę z obszarów o dużym zakrzywieniu i przenosi w okolice o mniejszej krzywiźnie. W miarę upływu czasu nasza powierzchnia coraz bardziej zaczyna przypominać powierzchnię o stałej dodatniej krzywiźnie, czyli euklidesową sferę. Mimo zmiany kształtu topologia się nie zmienia, zatem śledząc działanie przepływu Ricciego, możemy udowodnić, że wyjściowa powierzchnia w kształcie gruszki jest topologicznie równoważna sferze.

Ryc. 41. Jak przepływ Ricciego zmienia gruszkę w sferę

W tym przykładzie typ topologiczny naszej powierzchni był oczywisty od samego początku, ale taka sama ogólna strategia działa dla każdej rozmaitości. Możemy wyjść od dowolnie skomplikowanego kształtu i po prostu śledzić przepływ Ricciego. W miarę upływu czasu krzywizna będzie się rozkładała coraz bardziej równomiernie na wszystkich obszarach i analizowany kształt ulegnie uproszczeniu. Na samym końcu powinniśmy otrzymać najprostszy kształt o takiej samej topologii jak wyjściowa rozmaitość, bez

względu na to, jaka by ona była. W 1981 roku Hamilton udowodnił, że ta strategia sprawdza się w dwóch wymiarach, i tym samym przedstawił nowy dowód twierdzenia o klasyfikacji powierzchni.

Dokonał również znacznego postępu w stosowaniu analogicznego podejścia do rozmaitości trójwymiarowych, ale w tym wypadku pojawiła się poważna przeszkoda. W dwóch wymiarach każda powierzchnia automatycznie się upraszcza w wyniku działania przepływu Ricciego. Tak samo jest w trzech wymiarach, gdy wyjściowa rozmaitość ma w każdym punkcie ściśle dodatnią krzywiznę – nigdy zerową lub ujemną. Niestety, jeśli istnieją punkty, w których krzywizna jest zerowa – a często tak jest – to taka przestrzeń może się zaplątać w wyniku działania przepływu. Prowadzi to do powstania osobliwości – miejsc, w których rozmaitość przestaje być gładka. W takich punktach równania przepływu Ricciego ulegają załamaniu i proces rozdzielania krzywizny musi ulec zatrzymaniu. Naturalnym sposobem obejścia tej przeszkody jest zrozumienie, jak wyglądają takie rozmaitości, i odpowiednie ich przeprojektowanie – na przykład przez pocięcie ich na kawałki – tak by dało się uruchomić przepływ Ricciego. Jeśli dobrze zrozumiemy, jak topologia przebudowanej rozmaitości ma się do topologii rozmaitości wyjściowej, to takie podejście może się okazać skuteczne. Niestety Hamilton uświadomił sobie również, że w przypadku przestrzeni trójwymiarowych przepływ Ricciego może być naprawdę skomplikowany – najwyraźniej zbyt skomplikowany, by dało się wykorzystać tego rodzaju sztuczkę. Przepływ Ricciego szybko stał się standardową techniką stosowaną w geometrii, ale nie udało się za jego pomocą udowodnić hipotezy Poincarégo.

Minął rok 2000, a matematykom wciąż nie udało się uporać z tą hipotezą. Wówczas jednak jej wagę dodatkowo wzmocnił fakt, że uznano ją za jeden z siedmiu problemów milenijnych. Stało się też wtedy jasne, że gdyby ideę Hamiltona udało się w jakiś sposób zastosować w wystarczająco ogólnej sytuacji, to będzie z tego wynikała prawdziwość nie tylko hipotezy Poincarégo, ale także hipotezy geometryzacyjnej Thurstona. Nagroda kusiła wszystkich, ale uparcie pozostawała poza zasięgiem.

W matematyce jest tak samo jak we wszystkich pozostałych dziedzinach nauki – aby jakieś badania zostały uznane za poprawne, ich wyniki muszą zostać opublikowane, a to może nastąpić dopiero wtedy, gdy pracę zaakceptują recenzenci. Eksperci z danej dziedziny muszą przeczytać ją uważnie, sprawdzić logikę wywodu i upewnić się, że w obliczeniach nie ma błędów. W wypadku skomplikowanych i ważnych prac matematycznych proces ten może trwać dość długo. Jak wspomnieliśmy w rozdziale 1, kiedyś lekiem na tę bolączkę było rozsyłanie wstępnych wersji artykułu, teraz natomiast umieszcza się je na standardowym portalu internetowym arXiv (od ang. *archive* – archiwum), gdzie po przejściu uproszczonego procesu recenzji i akceptacji, którego celem jest wyłapanie oczywistych bzdur, przesłane wersje artykułu są udostępniane w postaci elektronicznej wszystkim zainteresowanym. Obecnie większość badaczy dowiaduje się o nowych wynikach właśnie z serwisu arXiv lub ze stron internetowych poszczególnych uczonych.

W 2002 roku do serwisu arXiv trafiła wersja wstępna artykułu Grigorija Perelmana poświęconego przepływowi Ricciego. Pojawiło się w nim niezwykłe stwierdzenie: przepływ ten ma charakter gradientowy. Innymi słowy, Perelman twierdził, że w przepływie Ricciego występuje dobrze określony kierunek „dół" i można zdefiniować określoną wielkość liczbową związaną z kształtem rozmaitości, która stale maleje z upływem czasu, gdy rozmaitość „przepływa w dół". Wielkość ta, którą można porównać do wysokości wzniesienia, pozwala w ilościowy sposób zmierzyć „uproszczenie" rozmaitości. Przepływy gradientowe mają dość ograniczone możliwości działania: nie mogą krążyć w koło ani zachowywać się chaotycznie. Wydaje się, że nikt nie podejrzewał, iż przepływ Ricciego może zostać aż tak ujarzmiony. Jednak Perelman nie ograniczył się tylko do wysunięcia takiego stwierdzenia – przedstawił również jego dowód. Na końcu artykułu zamieścił jeszcze szkic wywodu, który powinien pozwolić na udowodnienie hipotezy geometryzacyjnej Thurstona – z której wynika poprawność hipotezy Poincarégo i wiele znacznie głębszych wniosków – obiecując jednocześnie, że dalsze szczegóły prześle do serwisu arXiv

w kolejnych artykułach. W ciągu następnych ośmiu miesięcy przesłał dwa artykuły zawierające wiele z obiecanych szczegółów.

Pierwszy artykuł wywołał niemałe poruszenie. Perelman twierdził, że udało mu się wykonać cały program badawczy Hamiltona dzięki wykorzystaniu przepływu Ricciego do uproszczenia trójwymiarowej rozmaitości i udowodnieniu, że uzyskany wynik jest dokładnie taki, jak przewidział Thurston. Kolejne dwa artykuły umocniły przeświadczenie uczonych, że Perelman wie, o czym mówi i że jego pomysły są czymś więcej niż tylko szkicem obiecującej strategii z lukami logicznymi i bezpodstawnymi założeniami. Tym razem nie pojawiła się tradycyjna w takich wypadkach nieufność społeczności matematyków wobec wszelkich doniesień o rozwiązaniu wielkiego problemu. Zapanowało powszechne przeświadczenie, że Perelman faktycznie mógł tego dokonać.

Diabeł tkwi jednak w szczegółach, a w matematyce szczegóły potrafią być naprawdę diabelnie skomplikowane. Pracę Perelmana musieli sprawdzić wzdłuż i wszerz ludzie dobrze znający opisywane w niej gałęzie matematyki i świadomi wszystkich potencjalnych zagrożeń, jakie tam czyhają. Nie było to proste, ponieważ Perelman połączył ze sobą co najmniej cztery różne działy matematyki i fizyki matematycznej, a mało jest specjalistów, którzy znaliby więcej niż jeden lub dwa z nich. Ustalenie, czy dowód Perelmana jest poprawny, wymagało współpracy całego zespołu uczonych i dużego wysiłku z ich strony. Co więcej, artykuły przesłane do serwisu arXiv nie zawierały wszystkich szczegółów na poziomie, który zwykle spotyka się w pracach naukowych. Były napisane dość jasno, jak na wersje wstępne, ale Perelman nie zawsze rozpisywał wszystkie kroki. Eksperci musieli więc odtworzyć do pewnego stopnia tok myśli Perelmana – a on przecież zagłębiał się w tę tematykę już od wielu lat.

Wszystko to wymagało czasu. Perelman objaśniał swój dowód i odpowiadał na e-maile z pytaniami o różne kroki. Gdy tylko ktoś znajdował jakąś potencjalną lukę, szybko przesyłał dodatkowe wyjaśnienia, które rozwiewały wątpliwości. Sytuacja wyglądała zachęcająco. Jednak przed uzyskaniem całkowitej pewności, że w pracy nie ma żadnych błędów, nikt nie chciał ryzykować własnej

dobrej opinii i stwierdzić publicznie, że Perelman przedstawił dowód hipotezy Poincarégo, nie mówiąc już o trudniejszej hipotezie geometryzacyjnej. Zatem, mimo iż ogólnie wszyscy mieli bardzo dobrą opinię o pracy Perelmana, nikt się nie kwapił z oficjalnym uznaniem jej poprawności. Było to nieuniknione, ale doprowadziło niestety do nieprzyjemnej sytuacji, ponieważ w miarę jak przedłużało się oczekiwanie na opinię ekspertów, Perelmana coraz bardziej irytowało niezdecydowanie recenzentów. On *wiedział*, że dowód jest poprawny. Było to dla niego tak jasne, że nie potrafił pojąć, dlaczego inni mają trudności z jego zrozumieniem. Gdy go poproszono, żeby opisał swoją pracę bardziej szczegółowo lub przesłał artykuł do jakiegoś czasopisma – stanowczo odmówił. Z jego punktu widzenia sprawa była zamknięta, a w pracy przesłanej do serwisu arXiv znalazły się wszystkie potrzebne szczegóły. Przestał odpowiadać na pytania o rzekomo brakujące elementy wywodu. W jego opinii niczego tam nie brakowało. Na miłość boską, ludzie, chyba możecie poradzić sobie ze zrozumieniem tego bez dalszej pomocy z mojej strony. To nie takie trudne.

W niektórych relacjach z owych wydarzeń pojawiły się sugestie, że pod tym względem społeczność matematyków zachowała się nieuczciwie wobec Perelmana. Takie opinie świadczą jednak o braku zrozumienia tego, jak zachowują się matematycy, gdy ktoś twierdzi, że udało mu się rozwiązać jakiś wielki problem. Skrajnym przejawem nieodpowiedzialności byłoby w takiej sytuacji poklepywanie pretendenta po plecach i zapewnianie, że wykonał świetną robotę, a przymykanie oka na brakujące szczegóły w jego pracy. Zupełnie na miejscu, ba, nawet nieunikniona, była prośba o przygotowanie bardziej obszernego artykułu nadającego się do publikacji. W wypadku problemów takiej rangi pośpiech jest niebezpieczny i całkowicie nie do przyjęcia. W istocie eksperci zrobili więcej, niż się od nich oczekuje, ponieważ spędzili nad dowodem Perelmana wiele czasu i trzymali na wodzy swój naturalny sceptycyzm. Jeśli już chcemy mówić o wyjątkowym potraktowaniu Perelmana, to możemy najwyżej stwierdzić, że jego pracę przyjęto bardziej życzliwie, niż to się zwykle dzieje. W ostatecznym rozrachunku, gdy cały ten proces dobiegł końca, eksperci uznali, że jego dowód jest poprawny.

Wtedy jednak Perelman dawno stracił już cierpliwość. Zapewne nie bez znaczenia był fakt, że rozwiązał tak ważny problem, z którym nic nie może się równać. Był jak himalaista, który wspiął się na Mount Everest samotnie, bez korzystania z aparatu tlenowego. Nie ma już większego wyzwania. Czuł głęboką odrazę do szumu medialnego wokół jego dokonania – pragnął znaleźć uznanie w oczach kolegów, nie prezenterów telewizyjnych. Nie było więc zupełnym zaskoczeniem, że gdy w końcu inni matematycy przyznali, że ma rację, i postanowili uhonorować go Medalem Fieldsa i nagrodą Instytutu Claya, Perelman nie chciał o niczym słyszeć.

Dowód Perelmana, tak głęboki i elegancki, otwiera zupełnie nowy świat topologii. Perelman dokończył w nim program badawczy Hamiltona dotyczący wykorzystania przepływu Ricciego, wymyślając sprytne sposoby na poradzenie sobie z osobliwościami. Jednym z nich jest pozbycie się osobliwości przez przeskalowanie przestrzeni i czasu. W sytuacji gdy metoda ta zawodzi – mówimy wówczas, że osobliwość ulega kolapsowi – Perelman przeprowadza szczegółową analizę geometrii przepływu Ricciego, by ustalić, jak w danym wypadku dochodzi do kolapsu. Można powiedzieć, że przestrzeń wysuwa coraz cieńsze wypustki, czasami w dużej liczbie, przypominające gałęzie drzewa. Za każdym razem gdy taka wypustka jest o krok od kolapsu, można ją przeciąć, odcinając jej spiczasty, ostro zakrzywiony koniec, i zastąpić go łagodnie wygiętym zakończeniem. W wypadku niektórych z tych wypustek przepływ Ricciego zatrzymuje się w miejscu – jeśli tak się dzieje, należy takie miejsca zostawić w spokoju. Jeśli nie, przepływ Ricciego można uruchomić ponownie. Zatem niektóre wypustki mają gładkie zakończenia, a w innych przepływ został chwilowo przerwany, ale nadal działa.

Taka procedura, polegająca na obcinaniu końcówek i zastępowaniu ich łagodnymi zakończeniami, dzieli przestrzeń na fragmenty w zasadzie w taki sam sposób jak zaproponowane przez Thurstona cięcie przestrzeni na kawałki o jednorodnej geometrii. Okazuje się, że obie metody dają mniej więcej takie same wyniki. Jeden szczegół ma tu jednak kluczowe znaczenie: końcówki wymagające wstawienia gładkiego zakończenia nie pojawiają się coraz szybciej, dzięki czemu

nieskończenie dużą liczbę takich operacji można przeprowadzić w skończonym czasie. Jest to jedna z najtrudniejszych części dowodu.

Niektórzy komentatorzy krytykowali społeczność matematyków za niesprawiedliwe potraktowanie Perelmana. Nikt nie powinien być obojętny na słowa krytyki i faktycznie doszło do kilku incydentów, które można by uznać za przejaw niesprawiedliwości lub braku rozwagi, ale mimo to należy przyznać, że społeczność matematyków zareagowała na pracę Perelmana szybko i życzliwie. Reakcjom tym towarzyszyła jednak również ostrożność – rzecz całkowicie normalna i zrozumiała w matematyce i naukach ścisłych. Nieuniknione zainteresowanie środków masowego przekazu, dodatkowo podsycone perspektywą zdobycia wysokiej nagrody pieniężnej, wywarło wpływ na wszystkich, również na głównego bohatera.

Od czasu gdy w listopadzie 2002 roku Perelman zamieścił w serwisie arXiv swój pierwszy artykuł, do chwili, gdy w marcu 2010 roku przyznano mu nagrodę Instytutu Claya, upłynęło osiem lat. Mogłoby się wydawać, że to zbyt duża zwłoka. Jednak w pierwszym artykule Perelman poruszył tylko część problemu. Większość brakującego materiału zamieścił w serwisie w marcu 2003 roku. Do września 2004 roku, osiemnaście miesięcy po zamieszczeniu drugiego artykułu, specjaliści od przepływu Ricciego i topologii zdążyli przejrzeć dowód – a zabrali się do tego zaledwie kilka dni po pierwszym artykule – i najważniejsi eksperci oznajmili, że „rozumieją dowód". Znaleźli w nim pomyłki i luki, ale byli przekonani, że wszystkie da się wyjaśnić. Mówiąc prawdę, osiemnaście miesięcy to niezwykle krótko, gdy stawką jest coś tak ważnego.

Pod koniec 2005 roku przedstawiciele Międzynarodowej Unii Matematycznej powiadomili Perelmana, że na Międzynarodowym Kongresie Matematyków w 2006 roku chcieliby przyznać mu najwyższe wyróżnienie w tej dziedzinie – Medal Fieldsa. Kongres odbywa się co cztery lata, a więc była to pierwsza okazja, by wyrazić uznanie dla jego pracy. Ponieważ wciąż jeszcze było kilka wątpliwości odnośnie do pełnego dowodu hipotezy Poincarégo – wciąż

znajdowano nowe usterki – postanowiono, że oficjalnie medal zostanie przyznany za pogłębienie wiedzy na temat przepływu Ricciego, czyli za tę część artykułów, którą uznano już za bezbłędną.

Szczegółowe warunki przyznania nagrody milenijnej można znaleźć na stronie internetowej Instytutu Claya. Wymaga się zwłaszcza, by przedstawione rozwiązanie zostało opublikowane w recenzowanym czasopiśmie, a dwa lata później społeczność matematyków wciąż musi uznawać je za poprawne. Po tym okresie do pracy przystępuje specjalny komitet doradczy, który podejmuje decyzję o wysunięciu rekomendacji do przyznania nagrody lub nie. Perelman nie spełnił pierwszego warunku i wydawało się, że w ogóle nie ma takiego zamiaru. Jego zdaniem wersje wstępne artykułów zamieszczone w serwisie arXiv były zupełnie wystarczające. W tej sytuacji Instytut Claya postanowił odstąpić od egzekwowania wymogu publikacji w recenzowanym czasopiśmie i uznano, że rozpoczął się statutowy dwuletni okres oczekiwania, którego celem jest upewnienie się, że uczeni nie wykryją w pracy już żadnych błędów czy innego typu problemów. Okres ten dobiegł końca w 2008 roku i wtedy rozpoczęto dalsze działania, zgodnie z procedurami obowiązującymi w Instytucie, opracowanymi specjalnie po to, by uniknąć przedwczesnego przyznania nagrody.

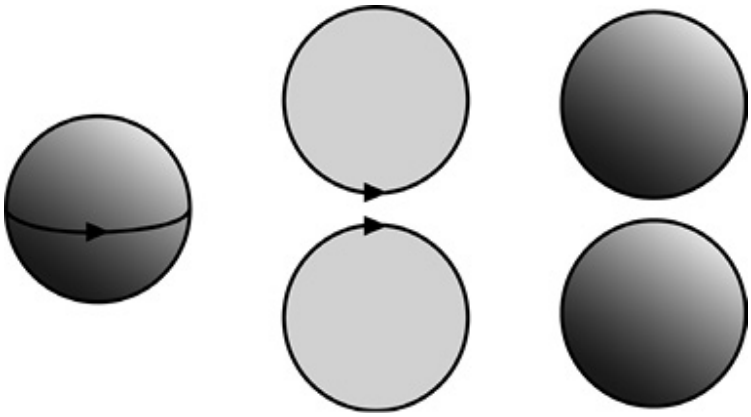
To prawda, że niektórzy eksperci ociągali się z przyznaniem, iż dowód jest poprawny. Powód tego był oczywisty: faktycznie nie mieli całkowitej pewności. Nie będzie zbyt dużą przesadą, jeśli stwierdzimy, że jedyną osobą zdolną szybko zrozumieć dowód Perelmana mógłby być tylko drugi Perelman. Doświadczony muzyk potrafi grać, odczytując z nut zupełnie nowy dla niego utwór, ale w wypadku dowodu matematycznego nie jest to możliwe. Matematycy muszą się najpierw przekonać, że wszystko w dowodzie jest sensowne. Wiedzą, że w każdym miejscu, w którym wywód robi się skomplikowany, istnieje duże prawdopodobieństwo popełnienia błędu. Tak samo jest zresztą wtedy, gdy idee stają się zbyt proste – wiele obiecujących dowodów poległo na założeniu tak oczywistym, że wydawało się, iż nie wymaga ono dowodu. Było więc zupełnie zrozumiałe, że opinię na temat przedstawionego dowodu można wyrazić dopiero wtedy, gdy eksperci nabiorą całkowitej pewności, że

jest on w zasadzie poprawny – i wtedy otwarcie uznali osiągnięcie Perelmana, *mimo* iż w dowodzie wciąż były luki i błędy. Pomyślmy tylko o całym tym zamieszaniu, jakie wywołały błędne, jak się w końcu okazało, prace na temat zimnej fuzji. W takich wypadkach wyrazem profesjonalnego podejścia jest rozwaga i postępowanie zgodnie z wyświechtanym frazesem: nadzwyczajne twierdzenia wymagają nadzwyczajnych dowodów.

Dlaczego Perelman odmówił przyjęcia Medalu Fieldsa i nagrody Instytutu Claya? Tylko on zna odpowiedź na to pytanie, ale sam wielokrotnie powtarzał, że nie interesują go tego rodzaju wyrazy uznania. Już wcześniej odmówił przyjęcia pomniejszych nagród. Od samego początku dał wszystkim jasno do zrozumienia, że nie chce zbyt wcześnie zyskać sławy. Jak na ironię, właśnie z tego samego, zrozumiałego powodu eksperci nie chcieli przedwcześnie wyrażać swej opinii. Patrząc na to realistycznie, należy stwierdzić, że nie było najmniejszych szans, by *uniknąć* przedostania się takiej informacji do mediów. Przez wiele lat matematycy robili, co mogli, żeby zainteresować swoją pracą prasę, radio i telewizję. Nie powinni się więc skarżyć, gdy ich wysiłki w końcu zaowocowały, ani oczekiwać, że dziennikarze przejdą do porządku dziennego nad najbardziej sensacyjną historią od czasu udowodnienia wielkiego twierdzenia Fermata. Perelman jednak tak tego nie postrzegał i zamknął się w swojej skorupie. Zaproponowano, że jeśli tylko się zgodzi, pieniądze z nagrody zostaną przeznaczone na edukację lub inne pożyteczne cele. Do tej pory nie odpowiedział.

74 Jednostkową trójwymiarową hipersferę tworzą wszystkie punkty o współrzędnych (*x*, *y*, *z*, *w*) takich, że $x^2 + y^2 + z^2 + w^2 = 1$. Istnieje kilka sposobów na oswojenie się z pojęciem trójwymiarowej hipersfery. Wszystkie można zrozumieć dzięki zastosowaniu analogii do dwuwymiarowej sfery i sprawdzić je za pomocą geometrii analitycznej. Jeden z takich opisów (kula, której całą powierzchnię utożsamiono z pojedynczym punktem) zamieściłem w tekście głównym, natomiast na rycinie 57 pokazałem jeszcze inne podejście. Zauważmy, że gdy przetniemy dwuwymiarową sferę wzdłuż równika, to dostaniemy dwie połówki wyglądające jak powierzchnie dwóch półkul. Każdą z tych połówek można przekształcić w sposób ciągły w koło. Aby odtworzyć dwuwymiarową sferę, musimy jedynie utożsamić ze sobą odpowiednie punkty na

brzegach obu kół. W pewnym sensie sporządziliśmy w ten sposób mapę dwuwymiarowej sfery na dwóch kołach, stosując proces podobny do tego, jakim posługują się kartografowie, przygotowując płaskie rzuty naszej okrągłej planety. W podobny sposób możemy skonstruować trójwymiarową hipersferę. W tym celu należy utożsamić ze sobą odpowiednie punkty na powierzchni dwóch kul. Taka operacja sprawia, że obie kule mają tę samą powierzchnię (ponieważ utożsamiliśmy ze sobą ich punkty) będącą dwuwymiarową sferą, która stanowi „równik" trójwymiarowej hipersfery.



Ryc. 57. Przepis na uzyskanie trójwymiarowej hipersfery. Po lewej: Rozcięcie dwuwymiarowej sfery na dwie połówki. W środku: Odtworzenie dwuwymiarowej sfery z dwóch połówek przez sklejenie ich brzegów. Po prawej: Podobnie możemy konceptualnie skleić powierzchnie dwóch kul i tym samym utożsamić ze sobą odpowiadające sobie punkty. W ten sposób otrzymamy trójwymiarową hipersferę.

75 W matematyce przyjęło się, że gdy mówimy o grupach, używamy dodawania i notacji $a + b$, jeśli obowiązuje prawo przemienności, a gdy rozważamy działanie nieprzemienne, mówimy o mnożeniu i używamy notacji ab. Nie zastosowałem się tutaj do tej konwencji, ponieważ nie jest to podręcznik teorii grup, a „dodawanie" wydaje się bardziej naturalne.

76 Zacznijmy liczyć od zera. Za każdym razem, gdy mijamy przystanek, jadąc w kierunku przeciwnym do ruchu wskazówek zegara, zwiększamy licznik o 1, a za każdym razem, gdy mijamy go, jadąc w drugą stronę, zmniejszamy licznik o 1. Po zakończeniu przejazdu dodajemy jeszcze 1, jeśli dojechaliśmy do przystanku, jadąc w kierunku przeciwnym do ruchu wskazówek zegara, jeśli natomiast

przyjechaliśmy z kierunku przeciwnego – odejmujemy 1. Tak uzyskany wynik jest całkowitą liczbą okrążeń mierzonych w kierunku przeciwnym do ruchu wskazówek zegara.

11. Nie wszystko może być proste

Problemy P i NP

W dzisiejszych czasach matematycy wykorzystują komputery na co dzień do rozwiązywania różnych zadań, nawet wielkich problemów. Komputery są dobre w arytmetyce, ale matematyka to coś więcej niż tylko obliczanie sum, zatem wprowadzenie interesującego problemu do komputera rzadko kiedy jest proste. Często najtrudniejszą częścią całego zadania jest przekształcenie go do postaci nadającej się do obliczenia na komputerze, ale nawet po takim przekształceniu komputery nierzadko mają duże kłopoty z realizacją zadania. W ostatnich czasach matematycy rozwiązali wiele ważnych problemów, ale w swoich pracach najczęściej nie stosowali komputerów albo wykorzystywali je tylko w niewielkim stopniu. Dobrym tego przykładem mogą być wielkie twierdzenie Fermata i hipoteza Poincarégo.

Gdy już używają komputerów do rozwiązania wielkiego problemu, na przykład twierdzenia o czterech barwach czy hipotezy Keplera, to odgrywają one w zasadzie rolę służebną. Czasami jednak role się odwracają i wtedy matematycy usługują informatyce. W większości przypadków zaprojektowanie pierwszych komputerów wymagało solidnej wiedzy matematycznej, przykładem tego może być oczywisty związek między algebrą Boole'a – algebraicznym ujęciem logiki – a układami przekaźników, takimi jak te, które projektował inżynier Claude Shannon, twórca teorii informacji. Obecnie zarówno praktyczne, jak i teoretyczne aspekty informatyki bazują na daleko

idącym wykorzystaniu rozwiązań matematycznych pochodzących z różnych dziedzin.

Jeden z problemów milenijnych wskazanych przez Instytut Claya leży na pograniczu matematyki i informatyki. Można go postrzegać dwojako: jako dowód na to, że informatyka odgrywa służebną rolę wobec matematyki, lub potwierdzenie przekonania, że to matematyka jest na służbie informatyki. W istocie jednak wymaga on bardziej zrównoważonego podejścia: partnerstwa – i pomaga w nawiązaniu takiej właśnie równorzędnej współpracy. Problem ten dotyczy algorytmów komputerowych, czyli szkieletów matematycznych, na podstawie których pisze się programy. Kluczowym pojęciem jest tu złożoność czasowa algorytmu, będąca miarą jego wydajności, która polega na ustaleniu, ile trzeba wykonać kroków obliczeniowych, by uzyskać odpowiedź przy określonym rozmiarze danych wejściowych. Mówiąc w uproszczeniu, wielkość ta mówi o tym, ile czasu potrzebuje komputer na rozwiązanie problemu o danym rozmiarze.

Słowo algorytm wywodzi się ze średniowiecza, gdy Muhammad ibn Musa al-Chuwarizmi napisał jedną z pierwszych książek poświęconych algebrze. Wcześniej Diofantos wprowadził pewien element, który powszechnie kojarzy się z algebrą: zapis symboliczny. Diofantos traktował jednak symbole tylko jako skróty i swoje metody rozwiązywania równań wyjaśniał, podając konkretne – choć typowe – przykłady. W miejscu, gdzie my zapisalibyśmy obecnie stwierdzenie typu „$x + a = y$, a zatem: $x = y - a$", Diofantos napisałby „załóżmy, że $x + 3 = 10$, a wtedy $x = 10 - 3 = 7$" i oczekiwał, że czytelnicy sami zrozumieją, iż takie podejście będzie nadal poprawne, jeśli zastąpimy 3 i 10 dowolnymi innymi liczbami. Wyjaśniając swoje przykłady, posługiwał się symbolami, ale nie przeprowadzał na nich żadnych działań. Al-Chuwarizmi opisywał ogólne metody w sposób jawny. Wykorzystywał do tego słowa, nie symbole, ale rozumiał podstawy takiego podejścia, i to właśnie jego uważa się powszechnie za ojca algebry. Nazwa tej dziedziny matematyki wywodzi się nawet z tytułu jednej z jego książek: *Al-kitāb al-mukhtaṣar fī ḥisāb al-ğabr wa'l-muqābala* (*Kompendium*

obliczeń za pomocą uzupełniania i porównywania). Słowo *al-ğabr* przekształciło się w „algebra”. Nazwa „algorytm” pochodzi natomiast od średniowiecznej, łacińskiej wersji jego nazwiska: Algorismus. Obecnie oznacza ona określoną procedurę matematyczną rozwiązywania jakiegoś zadania, która na pewno doprowadzi do uzyskania wyniku, jeśli się tylko wystarczająco długo poczeka.

W matematyce tradycyjnie przyjmuje się, że dany problem można w zasadzie uważać za rozwiązany, jeśli potrafimy zapisać algorytm prowadzący do uzyskania jego rozwiązania. Matematycy rzadko posługują się tym słowem, częściej mówią na przykład o wzorze na rozwiązanie, ale jest on jednak w istocie szczególnym rodzajem algorytmu zapisanego w języku symbolicznym. Nie jest zresztą ważne, czy wzór taki da się zastosować w praktyce – sam wzór *stanowi* rozwiązanie. Wszystko to zmieniło się za sprawą wykorzystania komputerów, ponieważ niektóre wzory, zbyt skomplikowane, by można je obliczać ręcznie, można teraz stosować w praktyce, angażując do tego komputery. Czasami jednak dochodzi do dość rozczarowującej sytuacji, gdy okazuje się, że dany wzór wciąż jest zbyt skomplikowany – można go wprawdzie wprowadzić do komputera i uruchomić obliczenia, ale działanie algorytmu jest zbyt wolne, by można było się doczekać odpowiedzi. Uczeni zaczęli więc szukać algorytmów wydajnych. Zarówno matematycy, jak i informatycy zainteresowali się opracowywaniem algorytmów, które naprawdę mogą gwarantować uzyskanie odpowiedzi w sensownym czasie.

Jeśli mamy konkretny algorytm, nietrudno ustalić, ile czasu (mierzonego liczbą wymaganych kroków obliczeniowych) zabierze znalezienie za jego pomocą rozwiązania dla danych wejściowych o ustalonym rozmiarze. Może to wymagać zastosowania odpowiednich technik, ale wiadomo, jak przebiega dany proces i co się w każdym momencie dzieje. Znacznie trudniejsze jest opracowanie bardziej wydajnego algorytmu, gdy jakiś znany algorytm okazuje się nieefektywny. A jeszcze trudniejsze jest ustalenie, jak dobry lub zły może być najwydajniejszy algorytm rozwiązujący dany problem, ponieważ wymaga to rozważenia

wszystkich możliwych algorytmów, a przecież nie mamy pojęcia, jak one wyglądają.

Pierwsze prace poświęcone tego typu zagadnieniom doprowadziły do wykształcenia się mało szczegółowego, ale wygodnego podziału algorytmów na dwie kategorie: na wydajne i niewydajne (w bardzo pragmatycznym znaczeniu tego słowa). Jeśli długość obliczeń rośnie względnie wolno w miarę wzrostu rozmiaru danych wejściowych, to taki algorytm jest wydajny i problem staje się łatwy do rozwiązania. Jeśli zaś długość obliczeń rośnie coraz szybciej ze wzrostem wielkości danych wejściowych, to taki algorytm jest niewydajny, a dany problem – trudny do rozwiązania. Z doświadczenia wiadomo, że choć niektóre problemy są łatwe w takim rozumieniu tego słowa, to jednak większość stanowią problemy trudne. Gdyby wszystkie problemy matematyczne były łatwe, matematycy dawno już straciliby pracę. Wspomniane na początku rozdziału wyzwanie, które znalazło się na liście problemów milenijnych, polega na przedstawieniu ścisłego dowodu na istnienie przynajmniej jednego problemu trudnego lub udowodnieniu – wbrew temu, co podpowiada nam doświadczenie – że wszystkie problemy są proste. Jest to tak zwany problem P/NP i nikt nie ma najmniejszego pojęcia, jak można by go rozstrzygnąć.

W rozdziale 2 poznaliśmy już jedną prostą miarę złożoności. Mówimy, że algorytm jest klasy P, jeśli czas jego wykonania ma charakter wielomianowy. Innymi słowy, w takim algorytmie liczba kroków, jakie trzeba wykonać, żeby otrzymać odpowiedź, jest proporcjonalna do jakiejś stałej potęgi, na przykład kwadratu lub sześcianu, wielkości danych wejściowych. Ogólnie rzecz biorąc, takie algorytmy są wydajne. Jeśli daną wejściową jest pojedyncza liczba, to jej rozmiarem jest liczba cyfr, a nie jej wartość. Wynika to stąd, że ilość informacji, jaka jest potrzebna do podania określonej liczby, jest równa ilości miejsca, jakie liczba ta zajmuje w pamięci komputera, a wielkość tego miejsca jest proporcjonalna do liczby cyfr. Mówimy, że problem jest klasy P, jeśli istnieje jakiś algorytm klasy P, który go rozwiązuje.

Wszystkie pozostałe algorytmy i problemy należą do klasy nie-P i większość z nich jest niewydajna. Są wśród nich i takie, których czas działania rośnie wykładniczo wraz ze wzrostem rozmiaru danych wejściowych – jest on w przybliżeniu równy jakiejś stałej wartości podniesionej do potęgi równej wielkości danych wejściowych. Są to algorytmy klasy E (od ang. *exponential* – wykładniczy) i uznaje się, że są niewydajne.

Niektóre natomiast są tak wydajne, że działają znacznie krócej, niż wynosi czas wielomianowy. Aby na przykład ustalić, czy dana liczba jest parzysta, wystarczy sprawdzić jej ostatnią cyfrę. Jeśli cyfrą tą jest (w notacji dziesiętnej) 0, 2, 4, 6, lub 8, to liczba jest parzysta, w przeciwnym razie jest nieparzysta. Taki algorytm wymaga wykonania najwyżej sześciu kroków:

Czy ostatnią cyfrą jest 0?
Jeśli tak, to KONIEC.
Liczba jest parzysta.

Czy ostatnią cyfrą jest 2?
Jeśli tak, to KONIEC.
Liczba jest parzysta.

Czy ostatnią cyfrą jest 4?
Jeśli tak, to KONIEC.
Liczba jest parzysta.

Czy ostatnią cyfrą jest 6?
Jeśli tak, to KONIEC.
Liczba jest parzysta.

Czy ostatnią cyfrą jest 8?
Jeśli tak, to KONIEC.
Liczba jest parzysta.

KONIEC. Liczba jest nieparzysta.

Zatem czas działania tego algorytmu jest równy najwyżej 6, niezależnie od wielkości danej wejściowej. Należy on do klasy o „stałym czasie”.

Ułożenie listy słów w porządku alfabetycznym jest problemem klasy P. Najprostszym sposobem wykonania tak postawionego zadania jest zastosowanie algorytmu sortowania bąbelkowego – nosi taką nazwę dlatego, że słowa znajdujące się na liście niżej, niż powinny być, przemieszczają się w górę niczym bąbelki w napoju gazowanym. Algorytm ten polega na wielokrotnym przeglądaniu listy, porównywaniu sąsiednich słów i zamienianiu ich miejscami, jeśli są w złej kolejności. Przypuśćmy na przykład, że nasza lista wygląda na początku następująco:

WÓŁ RYŚ KOT BYK

Po pierwszym przejściu zmienia się następująco:

RYŚ WÓŁ KOT BYK
RYŚ **KOT WÓŁ** BYK
RYŚ KOT **BYK WÓŁ**

przy czym tłustym drukiem zaznaczono wyrazy porównywane w danym kroku. Przy drugim przebiegu lista zmienia się następująco:

KOT RYŚ BYK WÓŁ
KOT **BYK RYŚ** WÓŁ
KOT BYK **RYŚ WÓŁ**

Trzeci przebieg daje następujący wynik:

BYK KOT RYŚ WÓŁ
BYK **KOT RYŚ** WÓŁ
BYK KOT **RYŚ WÓŁ**

Przy czwartym przebiegu żaden z wyrazów nie zmieni już miejsca i tym samym wiemy, że wykonaliśmy zadanie. Zwróćmy uwagę na to, jak słowo BYK przesuwa się powoli w górę (a raczej do przodu) listy.

Przy czterech słowach nasz algorytm wykonuje trzy operacje porównania w każdym z czterech przebiegów. W wypadku listy zawierającej *n* słów będzie to *n* – 1 porównań na przebieg i *n* przebiegów, a więc w sumie $n(n-1)$ kroków. Sortowanie bąbelkowe okazuje się łatwym algorytmem klasy P, ale jego wydajność nie jest nawet zbliżona do wydajności najlepszych algorytmów sortujących. Najszybszy algorytm sortowania z porównywaniem, zaprojektowany w sposób dużo bardziej sprytny, wymaga wykonania $n \ln n$ kroków.

Prostym algorytmem o wykładniczym czasie działania, czyli algorytmem klasy E, jest algorytm wypisujący listę wszystkich liczb dwójkowych (binarnych) o *n* cyfrach. Taka lista zawiera 2^n liczb i wypisanie (oraz obliczenie) każdej z nich zabiera w przybliżeniu *n* kroków, a zatem czas działania algorytmu wynosi około $2^n n$. To więcej niż 2^n, ale mniej niż 3^n, gdy *n* jest odpowiednio duże. Jest to jednak dość głupi przykład, ponieważ algorytm ten jest powolny z uwagi na rozmiar danych wyjściowych, a nie ze względu na wielkość danych wejściowych czy złożoność obliczeniową – to rozróżnienie okaże się kluczowe w dalszej części naszych rozważań.

Bardziej typowy algorytm klasy E rozwiązuje problem komiwojażera. Pewien komiwojażer musi odwiedzić określoną liczbę miast. Może do nich pojechać w dowolnej kolejności. Która z możliwych tras jest najkrótsza? Najbardziej bezpośrednia metoda rozwiązania tak postawionego zadania polega na sporządzeniu listy wszystkich możliwych tras, wyznaczeniu całkowitej długości każdej z nich i znalezieniu najkrótszej. Dla *n* miast mamy

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

tras (zapis *n*! czytamy „n silnia”). Taka złożoność rośnie szybciej niż dowolna złożoność wykładnicza[77]. Bardziej wydajna metoda, nazywana programowaniem dynamicznym, pozwala rozwiązać problem komiwojażera w czasie wykładniczym. Pierwsze tego typu rozwiązanie, tak zwany algorytm Helda–Karpa, znajduje najkrótszą

trasę w $2^n n^2$ krokach – wartość ta mieści się między 2^n i 3^n, gdy *n* jest odpowiednio duże.

Mimo że algorytmy te są „niewydajne”, można zastosować specjalne sztuczki pozwalające skrócić czas obliczeń, gdy liczba miast jest duża według naszych ludzkich standardów, ale nie na tyle duża, by takie sztuczki przestały działać. W 2006 roku David L. Applegate, Robert M. Bixby, Václav Chvátal i William J. Cook rozwiązali problem komiwojażera dla 85 900 miast – w połowie 2012 roku, gdy piszę te słowa, jest to wciąż najlepszy wynik[78].

Te przykłady algorytmów służą nie tylko zilustrowaniu pojęcia wydajności. Potwierdzają również moją wcześniejszą uwagę o tym, jak trudno jest znaleźć lepsze wersje znanych algorytmów, nie mówiąc już o algorytmach o największej możliwej wydajności. Wszystkie znane algorytmy rozwiązujące problem komiwojażera są klasy E, czyli mają wykładniczy czas wykonania – nie oznacza to jednak, że nie mogą istnieć algorytmy bardziej wydajne. Świadczy jedynie o tym, że takich algorytmów jak dotąd nie znaleziono. Są dwie możliwości: albo nie odkryliśmy lepszego algorytmu, ponieważ nie jesteśmy wystarczająco sprytni, albo dlatego, że taki nie istnieje.

Dobry tego przykład przedstawiliśmy w rozdziale 2. Zanim zespół Agrawala znalazł algorytm klasy P na test pierwszości, najlepszy znany algorytm tego rodzaju był klasy nie-P. Wydawał się całkiem niezły – czas jego wykonania wynosił $n \ln n$ dla *n* cyfrowych liczb, co oznacza, że jest nawet lepszy od algorytmu AKS (Agrawala, Kayala i Saxeny) dla wszystkich liczb o długości mniejszej od 10^{1000} cyfr. Przed odkryciem algorytmu AKS opinia specjalistów na temat klasy tego problemu była podzielona. Niektórzy eksperci podejrzewali, że test pierwszości jest problemem klasy P i kiedyś w końcu uda się znaleźć odpowiedni algorytm, inni natomiast nie zgadzali się z tą opinią. Nowy algorytm pojawił się jak grom z jasnego nieba – był jednym z tysięcy różnych pomysłów, które można sprawdzić, i przypadkiem okazało się, że ten akurat działa. To niezwykle wymowny fakt świadczący o tym, że nie znamy odpowiedzi na tak

postawione pytania i opinie ekspertów mogą być zgodne z prawdą, ale wcale nie muszą.

Rozważany w tym rozdziale problem milenijny związany jest jednak z odpowiedzią na bardziej podstawowe pytanie. Czy w ogóle istnieją jakieś trudne problemy? A może wszystkie są proste i musimy się tylko wykazać odpowiednim sprytem, by to stwierdzić? W rzeczywistości zagadnienie to trzeba sformułować nieco precyzyjniej, ponieważ spotkaliśmy się już z przykładem problemu, który bez wątpienia jest trudny – chodzi o przykład z wypisaniem wszystkich *n* cyfrowych liczb dwójkowych. Jak wspomnieliśmy, przykład ten był niezbyt trafny – cała trudność nie zawiera się w nim w obliczeniach, ale w konieczności wypisania bardzo długiej odpowiedzi. Wiemy, że nie ma możliwości znalezienia szybszego algorytmu, ponieważ odpowiedź ma taką długość *z definicji*. Gdyby była krótsza, nie byłaby rozwiązaniem tak postawionego zadania.

Aby nasze pytanie miało sens, musimy wykluczyć tego typu trywialne przykłady. W tym celu należy wprowadzić nową klasę algorytmów – klasę NP. Nie jest to skrót od nie-P. Do tej klasy należą algorytmy o niedeterministycznie wielomianowym czasie działania (skrót NP pochodzi od ang. *nondeterministic polynomial*). To fachowe określenie oznacza, że bez względu na to, ile czasu potrzebuje dany algorytm na znalezienie rozwiązania, zawsze możemy *sprawdzić jego poprawność* w czasie wielomianowym. Znalezienie odpowiedzi może być trudne, ale gdy ją już mamy, potrafimy w łatwy sposób upewnić się, że jest poprawna.

W nazwie tej klasy używa się przymiotnika „niedeterministyczny”, ponieważ problem klasy NP można rozwiązać, próbując odgadnąć odpowiedź. Jeśli wytypujemy jakieś możliwe rozwiązanie, możemy sprawdzić, czy faktycznie jest ono poprawne. Jeśli na przykład zadanie polega na rozłożeniu na czynniki pierwsze liczby 11 111 111 111, możemy wysunąć przypuszczenie, że jednym z jej czynników jest liczba pierwsza 21 649. Podana tak po prostu, liczba ta wygląda na zupełnie przypadkową. Łatwo możemy jednak sprawdzić, czy jest czynnikiem liczby z zadania: wystarczy wykonać odpowiednie dzielenie. Okazuje się, że liczby te dzielą się bez reszty i wynikiem jest 513 239 – a więc udało nam się zgadnąć. Gdybyśmy

zaproponowali liczbę 21 647 – która również jest liczbą pierwszą – to w wyniku dzielenia dostalibyśmy 513 286 i 9069 reszty, a więc byłaby to nieudana próba odgadnięcia odpowiedzi.

Odgadnięcie poprawnej odpowiedzi wymaga w tym przykładzie chyba cudu lub zastosowania jakiejś sztuczki (ja na przykład wyznaczyłem czynniki liczby 11 111 111 111, zanim spróbowałem „zgadnąć"). O to jednak nam przecież chodzi. Gdyby nie było potrzeba cudu, to zgadując wielokrotnie aż do uzyskania poprawnej odpowiedzi, moglibyśmy zmienić algorytm klasy NP w algorytm klasy P. Nasz przykład pokazuje, dlaczego takie podejście nie zadziała: musielibyśmy zgadywać zbyt wiele razy. W istocie takie podejście jest równoważne z dzieleniem po kolei przez wszystkie możliwe liczby pierwsze tak długo, aż znajdziemy jeden z czynników. Z rozdziału 2 wiemy już, że jest to kiepska metoda szukania czynników pierwszych.

Wprowadzając klasę NP, możemy wykluczyć takie przypadki jak nasz przykład z bardzo długą listą. Jeśli odgadniemy listę wszystkich *n* cyfrowych liczb binarnych, to nie tylko ich wypisanie będzie wymagało czasu wykładniczego. Również sprawdzenie zaproponowanego rozwiązania będzie wymagało czasu wykładniczego, zatem zweryfikowanie odpowiedzi potrwa dłużej niż jej znalezienie. Polecenie sprawdzenia poprawności takiego rozwiązania mogłoby być koszmarem nawet dla najbardziej doświadczonego redaktora zajmującego się korektą. Nie ma wątpliwości, że klasa P zawiera się w klasie NP. Jeśli potrafimy znaleźć odpowiedź w czasie wielomianowym, mając przy tym gwarancję, że jest ona poprawna, to jest to równoważne ze sprawdzeniem jej poprawności. Zatem automatyczna weryfikacja wyniku wymaga w najgorszym wypadku czasu wielomianowego. Gdyby ktoś przedstawił nam jakąś proponowaną odpowiedź, moglibyśmy po prostu uruchomić jeszcze raz cały algorytm. To byłoby nasze sprawdzenie.

Teraz już możemy sformułować nasz problem milenijny. Czy klasa NP jest większa od P, czy też są takie same? Lub krócej: czy P jest równe NP?

Jeśli odpowiedź brzmi: „tak”, to powinniśmy móc znaleźć szybkie, wydajne algorytmy planowania harmonogramu lotów, optymalizacji produkcji zakładów przemysłowych czy wykonywania milionów innych ważnych zadań. Jeśli natomiast odpowiedź brzmi: „nie”, to będziemy mieli stuprocentową gwarancję, że wszystkie problemy, które obecnie sprawiają wrażenie trudnych, faktycznie takie są i w związku z tym będziemy mogli z czystym sumieniem przestać marnować czas na szukanie szybkich algorytmów realizujących takie zadania. Tak czy inaczej wygramy. Cały kłopot w tym, że nie mamy pojęcia, jak jest.

Życie matematyków stałoby się prostsze, gdyby odpowiedź brzmiała: „tak”, ale drzemiący w każdym z nas pesymista od razu podejrzewa jednak, że życie nie jest takie proste i odpowiedzią jest zapewne: „nie”. W przeciwnym razie dostalibyśmy za darmo coś, na co nie zapracowaliśmy ani nie zasłużyliśmy. Podejrzewam, że tak naprawdę większość matematyków wolałaby, żeby odpowiedź brzmiała: „nie”, ponieważ dzięki temu mieliby zapewnione zajęcie po wsze czasy. Matematycy dowodzą swojej wartości, rozwiązując trudne problemy. Z takich czy innych powodów większość matematyków i informatyków spodziewa się, że odpowiedź na pytanie „Czy P jest równe NP?” brzmi: „nie”. Niemal nikt nie zakłada, że może być inaczej.

Istnieją jeszcze dwie inne możliwości. Możemy na przykład znaleźć dowód, że P jest równe NP, nie odkrywszy jednak żadnego algorytmu o czasie wielomianowym dla określonych problemów NP. W matematyce często się zdarza, że znajdujemy dowody istnienia czegoś, z których nie wynikają żadne bezpośrednie konsekwencje. Taki dowód pokazuje, że coś istnieje, ale nie wynika z niego dokładnie, co to jest. Przykładem mogą być testy pierwszości, które radośnie informują nas, że dana liczba nie jest pierwsza, nie podając przy tym żadnego konkretnego czynnika pierwszego, lub twierdzenia z teorii liczb zapewniające, że rozwiązania danego równania diofantycznego są ograniczone – mniejsze od jakiejś wartości granicznej – nieprecyzujące żadnej określonej granicy. Algorytm o czasie wielomianowym mógłby być tak skomplikowany, że jego zapisanie byłoby po prostu niemożliwe. Wówczas nasz wrodzony

pesymizm wobec darmowych okazji byłby uzasadniony, nawet gdyby odpowiedź na główne pytanie była twierdząca.

Może być jeszcze gorzej – niektórzy badacze przypuszczają, że na tak postawione pytanie nie uda się jednoznacznie odpowiedzieć w ramach obecnego formalizmu logicznego obowiązującego w matematyce. W takim razie żadnej z odpowiedzi nie da się udowodnić. Nie dlatego, że nie jesteśmy dość sprytni, by znaleźć taki dowód, ale dlatego, że go nie ma. Taka możliwość stała się jasna w 1931 roku, gdy Kurt Gödel wpuścił kota nieokreśloności między filozoficzne myszy, które rozpleniły się w podstawach matematyki. Udowodnił w ten sposób, że pewne stwierdzenia arytmetyczne są niemożliwe do rozstrzygnięcia. W 1936 roku Alan Turing znalazł prostszy problem niemożliwy do rozstrzygnięcia – problem stopu maszyny Turinga. Czy dla danego algorytmu zawsze można znaleźć dowód, że się kiedyś zatrzyma, albo dowód, że będzie się wykonywał wiecznie? To zaskakujące, ale Turing pokazał, że odpowiedź brzmi: „nie". Dla niektórych algorytmów nie można przedstawić ani jednego, ani drugiego dowodu. Być może właśnie taki jest problem P/NP. Wyjaśniałoby to, dlaczego nikomu nie udało się go ani udowodnić, ani mu zaprzeczyć. Nikomu też nie udało się jednak dowieść ani zaprzeczyć, że problem P/NP jest nierozstrzygalny. Może jego nierozstrzygalność jest nierozstrzygalna…

Najbardziej bezpośrednim sposobem uporania się z problemem P/NP byłoby znalezienie zagadnienia, o którym wiadomo, że jest klasy NP, a następnie założenie, że istnieje jakiś algorytm rozwiązujący to zagadnienie w czasie wielomianowym, i doprowadzenie całej rzeczy do sprzeczności. Przez pewien okres uczeni próbowali zastosować taką metodę do wielu różnych problemów, ale w 1971 roku Stephen Cook uświadomił sobie, że wybór problemu często nie ma znaczenia. W pewnym sensie wszystkie tego typu problemy – z dokładnością do kilku szczegółów – są niczym muszkieterowie: jeden za wszystkie, wszystkie za jednego. Cook wprowadził pojęcie problemu NP-zupełnego. Jest to konkretny problem klasy NP cechujący się tym, że jeśli uda się znaleźć jakiś algorytm klasy P, który go rozwiązuje, to

będzie to oznaczało, iż *wszystkie* problemy klasy NP można rozwiązać za pomocą algorytmów klasy P.

Cook znalazł kilka problemów NP-zupełnych, między innymi tak zwany problem SAT, czyli problem spełnialności formuł z zakresu algebry Boole'a. Chodzi w nim o odpowiedź na pytanie, czy dane wyrażenie logiczne może być prawdziwe, jeśli odpowiednio przypiszemy do jego zmiennych wartości „prawda" i „fałsz". Cook zaproponował również znacznie głębszy przykład – pokazał mianowicie, że bardziej restrykcyjny problem, nazywany problemem 3-SAT, też jest problemem NP-zupełnym. W tym wypadku rozważane wyrażenie logiczne można zapisać w postaci: „*A* lub *B* lub *C* lub … lub *Z*", gdzie *A*, *B*, *C*, …, *Z* to wyrażenia logiczne zaledwie trzech zmiennych – dodajmy, że nie muszą to być za każdym razem te same trzy zmienne. Większość dowodów tego, że dany problem jest NP-zupełny, nawiązuje do twierdzenia Cooka na temat problemu 3-SAT.

Z definicji Cooka wynika, że wszystkie problemy NP-zupełne mają taką samą wagę. Jeśli udowodnimy, że jeden z nich jest klasy P, to będzie to oznaczało, że wszystkie są klasy P. Otwiera to przed nami możliwość zastosowania nowej taktyki: niektóre problemy NP-zupełne mogą był łatwiejsze od innych. Możemy więc wybrać sobie dowolny problem NP-zupełny i zająć się tylko nim – ze strategicznego punktu widzenia nie ma znaczenia, który z nich będziemy rozważać. Wszystkie problemy NP-zupełne są pod tym względem równorzędne, ponieważ każdy taki problem może udawać dowolny inny problem klasy NP. Dowolny problem klasy NP można przekształcić w specjalny przypadek problemu NP-zupełnego przez odpowiednie „przekodowanie", które daje się przeprowadzić w czasie wielomianowym.

Aby się przekonać, jak wygląda taka procedura, rozważmy typowy problem NP-zupełny polegający na znalezieniu cyklu Hamiltona w sieci. Chodzi o wytyczenie zamkniętej drogi przebiegającej przez krawędzie sieci w taki sposób, by przez każdy wierzchołek (punkt) przechodziła dokładnie raz. Droga zamknięta to taka, która wraca do punktu wyjścia. Rozmiar danych wejściowych jest w tym wypadku równy liczbie krawędzi, a ta jest mniejsza lub równa kwadratowi

liczby wierzchołków, ponieważ każda krawędź łączy dwa wierzchołki. (Zakładamy, że daną parę wierzchołków łączy najwyżej jedna krawędź). Nie znamy żadnego algorytmu klasy P, który pozwalałby rozwiązać tak postawiony problem, ale załóżmy hipotetycznie, że taki algorytm istnieje. Wybierzmy teraz jakiś inny problem i nazwijmy go problemem X. Przypuśćmy, że można go sprowadzić do równoważnego problemu znalezienia cyklu Hamiltona w jakiejś związanej z tym problemem sieci. Jeśli procedurę przekształcania danych opisujących problem X do postaci sieci i odwrotnie można wykonać w czasie wielomianowym, to automatycznie otrzymujemy algorytm klasy P dla problemu X. Wygląda on tak:

1. Przekształć X do postaci problemu znalezienia cyklu Hamiltona w sieci związanej z problemem X – tę operację można przeprowadzić w czasie wielomianowym.
2. Znajdź wymagany cykl w czasie wielomianowym, wykorzystując do tego hipotetyczny algorytm do rozwiązywania tego typu problemów.
3. Przekształć otrzymany cykl Hamiltona z powrotem w rozwiązanie problemu X – tę operację również można przeprowadzić w czasie wielomianowym.

Ponieważ trzy kolejne kroki zajmujące czas wielomianowy wykonują się w sumie również w czasie wielomianowym, taki algorytm jest algorytmem klasy P.

Aby pokazać, jak to działa, rozważymy mniej ambitną wersję problemu poszukiwania cyklu Hamiltona, w której nie wymaga się, by znaleziona droga była zamknięta. Jest to tak zwany problem drogi Hamiltona. W sieci może istnieć droga Hamiltona nawet wtedy, gdy nie ma w niej cyklu (zob. ryc. 42, po lewej), a więc rozwiązanie problemu cyklu Hamiltona nie musi być rozwiązaniem problemu drogi Hamiltona. Możemy jednak przekształcić problem drogi Hamiltona w problem cyklu Hamiltona, wykorzystując do tego zbliżoną, ale inną sieć. W tym celu należy dodać jeszcze jeden węzeł połączony ze wszystkimi węzłami pierwotnej sieci, tak jak to

pokazano na rycinie 42 (po prawej). Dowolny cykl Hamiltona w nowej sieci można przekształcić w drogę Hamiltona w sieci pierwotnej – wystarczy pominąć nowy węzeł i dochodzące do niego krawędzie cyklu. I odwrotnie: każda ścieżka Hamiltona w pierwotnej sieci prowadzi do powstania cyklu Hamiltona w sieci zmodyfikowanej – wystarczy połączyć oba końce drogi Hamiltona z dostawionym węzłem. Takie „przekodowanie" problemu drogi na problem cyklu wymaga wprowadzenia tylko jednego nowego węzła i po jednej nowej krawędzi na każdy węzeł w sieci pierwotnej. Zatem taką procedurę – i procedurę odwrotną – można wykonać w czasie wielomianowym.

Ryc. 42. Po lewej: Sieć z drogą Hamiltona (zaznaczoną czarnymi liniami), w której nie ma cyklu Hamiltona. Po prawej: Po wprowadzeniu dodatkowego węzła (zaznaczonego szarym kolorem) i czterech nowych krawędzi możemy przekształcić drogę Hamiltona w cykl (zaznaczony czarnymi liniami). Dwie krawędzie narysowane szarym kolorem nie stanowią części cyklu, ale są potrzebne do skonstruowania zmodyfikowanej sieci.

Oczywiście w tym wypadku ograniczyliśmy się tylko do przekodowania jednego określonego problemu do postaci problemu cyklu Hamiltona. Aby udowodnić, że problem cyklu Hamiltona jest NP-zupełny, należałby zdefiniować taką operację dla dowolnego problemu klasy NP. Można tego dokonać – jako pierwszy dowód taki znalazł Richard Karp w 1972 roku i zamieścił go w słynnym artykule, w którym udowodnił, że 21 różnych problemów należy do klasy problemów NP-zupełnych[79].

Problem komiwojażera jest „niemal" problemem NP-zupełnym – istnieje pewien formalny szkopuł: nie wiadomo, czy jest to problem

klasy NP. Obecnie znamy ponad 300 problemów NP-zupełnych z różnych dziedzin matematyki: logiki, sieci, kombinatoryki czy optymalizacji. Jeśli uda się udowodnić, że chociaż jeden z nich daje się (lub nie daje się) rozwiązać w czasie wielomianowym, będzie to równoznaczne z udowodnieniem tego dla wszystkich tych problemów. Mimo tej klęski urodzaju problem P/NP pozostaje kwestią otwartą. Nie zdziwiłbym się zbytnio, gdyby za sto lat okazało się, że wciąż nie potrafimy go rozwiązać.

77 Ze wzoru Stirlinga wynika, że *n*! jest równe w przybliżeniu

$$\sqrt{2\pi n}\left(n/e\right)^n$$

78 William J. Cook, *In Pursuit of the Travelling Salesman* (*W pogoni za komiwojażerem*), Princeton University Press, 2012. Bieżące informacje (w języku angielskim) można znaleźć na stronie internetowej http://www.math.uwaterloo.ca/tsp/index.html.

79 Richard M. Karp, *Reducibility among combinatorial problems* (*Redukowalność problemów kombinatorycznych*) [w:] R.E. Miller [red.], J.W. Thatcher [red.], *Complexity of Computer Computations* (*Złożoność obliczeń komputerowych*), Plenum, 1972, s. 85–103.

12. Rozmyślania o cieczach

Równania Naviera–Stokesa

Pięć spośród siedmiu problemów milenijnych, włącznie z trzema, które dotąd omówiliśmy, ma swoje źródło w matematyce teoretycznej, choć problem P/NP ma również podstawowe znaczenie dla informatyki. Pozostałe dwa dotyczą klasycznej matematyki stosowanej i współczesnej fizyki matematycznej. Problem milenijny z zakresu matematyki stosowanej dotyczy standardowego równania cieczy, które nazywamy równaniem Naviera–Stokesa od nazwisk francuskiego inżyniera i fizyka Claude'a Louisa Naviera i irlandzkiego matematyka i fizyka George'a Stokesa. Jest to równanie różniczkowe cząstkowe, co oznacza, że opisuje ono szybkość zmian przepływu cieczy zarówno w przestrzeni, jak i w czasie. Większość wielkich równań klasycznej matematyki stosowanej i fizyki również ma postać równań różniczkowych cząstkowych – w dalszej części książki spotkamy jeszcze jedno takie równanie: równanie Laplace'a – a nie zwyczajnych równań różniczkowych, opisujących szybkość zmian jakiejś wielkości wyłącznie względem czasu.

W rozdziale 8 mówiliśmy o tym, że ruch obiektów Układu Słonecznego wynika z prawa powszechnego ciążenia Newtona i jego zasad dynamiki. Prawa te wiążą przyspieszenie Słońca, Księżyca i planet z działającymi na nie siłami grawitacyjnymi. Przyspieszenie jest szybkością zmiany prędkości w czasie, a prędkość jest szybkością zmiany położenia w czasie. Jest to więc zwyczajne równanie różniczkowe. Jak się przekonaliśmy, równania takie mogą być bardzo trudne do rozwiązania. Rozwiązywanie cząstkowych równań różniczkowych jest zazwyczaj jeszcze trudniejsze.

W celu zastosowań praktycznych rozwiązania opisujące Układ Słoneczny można rozwiązać numerycznie na komputerze. To wciąż nie jest proste, ale udało się opracować dobre metody pozwalające przeprowadzać takie obliczenia. Podobnie sytuacja wygląda w przypadku praktycznych zastosowań równań Naviera–Stokesa. Stosuje się tu metody z dziedziny nazywanej obliczeniową (lub numeryczną) mechaniką płynów, które wykorzystuje się w bardzo różnorodnych dziedzinach szeroko rozumianej techniki: w projektowaniu samolotów, przy badaniu aerodynamiki karoserii samochodowych, a nawet w rozwiązywaniu problemów medycznych, takich jak przepływ krwi w ludzkim organizmie.

Autorzy listy problemów milenijnych nie wymagają od matematyków znalezienia jawnych rozwiązań równania Naviera–Stokesa, ponieważ jest to w zasadzie niemożliwe. Nie chodzi im też o przedstawienie metod numerycznych rozwiązywania tych równań, choć niewątpliwie są one ważne. Proszą jedynie o przedstawienie dowodu pewnej podstawowej własności teoretycznej, a mianowicie: *istnienia* rozwiązań. Czy mając dany stan cieczy w określonej chwili – wzorzec jej ruchu – można wyznaczyć rozwiązanie równania Naviera–Stokesa, które będzie prawdziwe dla każdego późniejszego czasu, począwszy od podanego stanu początkowego? Intuicja fizyczna podpowiada nam, że odpowiedź na tak postawione pytanie powinna brzmieć: „tak”, ponieważ równanie to jest bardzo dokładnym modelem fizycznym prawdziwych cieczy. Jednak z matematycznego punktu widzenia kwestia istnienia rozwiązania nie jest już tak jednoznaczna i takiej podstawowej własności tego równania nigdy nie udało się udowodnić. Może się więc jeszcze okazać, że nie jest to prawdą.

Równanie Naviera–Stokesa opisuje, jak w danych okolicznościach rozkład prędkości cieczy zmienia się z upływem czasu. Mówiąc o tym równaniu, często używamy liczby mnogiej – równania Naviera–Stokesa – ale w obu wypadkach chodzi o to samo. Liczba mnoga wynika z klasycznego podejścia: w trójwymiarowej przestrzeni prędkość ma trzy składowe i w tradycyjnym ujęciu każdą z nich

opisuje oddzielne równanie, a więc w sumie są trzy. W nowoczesnym ujęciu zapisuje się tylko jedno równanie dla *wektora* prędkości (wielkości mającej zarówno wartość, jak i kierunek), które można zastosować do każdej z trzech składowych prędkości. Na stronie internetowej Instytutu Claya używana jest tradycyjna terminologia, ale my będziemy się posługiwali nowoczesnym ujęciem. Wspominam tu o tym, żeby uniknąć ewentualnych nieporozumień.

Historia tego równania zaczęła się w 1822 roku, gdy Navier zapisał równanie różniczkowe cząstkowe przepływu cieczy lepkiej. Stokes wniósł swój wkład w latach 1842 i 1843. Euler zapisał równanie różniczkowe cząstkowe dla cieczy o zerowej lepkości już w 1757 roku. Chociaż to równanie wciąż jest użyteczne, większość rzeczywistych cieczy, w tym woda i powietrze, ma pewną lepkość, dlatego Navier i Stokes poprawili równanie Eulera, tak by uwzględnić ten fakt. Obaj uczeni wyprowadzili niezależnie w zasadzie to samo równanie, dlatego dzisiaj w jego nazwie występują dwa nazwiska. Navier popełnił parę błędów rachunkowych, ale uzyskał poprawne rozwiązanie. Stokes przeprowadził obliczenia bezbłędnie i stąd wiemy, że rozwiązanie Naviera było poprawne mimo pomyłek. W swojej najogólniejszej postaci równanie to stosuje się do cieczy ściśliwych, takich jak powietrze. Jest jednak pewien ważny przypadek szczególny, w którym zakłada się, że ciecz jest nieściśliwa. Taki model stosuje się do cieczy takich jak woda, które wprawdzie zmniejszają objętość pod wpływem działania dużych sił, ale nieznacznie.

Istnieją dwa sposoby przedstawienia matematycznego opisu przepływu cieczy: możemy opisać, jak przebiega w czasie droga każdej cząsteczki cieczy, lub scharakteryzować prędkość przepływu w każdym punkcie przestrzeni i w każdym momencie. Oba opisy są ze sobą związane: jeśli mamy jeden, to możemy – wkładając w to nieco wysiłku – wyprowadzić drugi. Euler, Navier i Stokes posłużyli się tym drugim, ponieważ prowadzi on do uzyskania równania łatwiejszego z matematycznego punktu widzenia. W swoich równaniach odwołują się więc do pola prędkości cieczy. W każdym określonym momencie pole prędkości określa szybkość i kierunek ruchu każdej cząsteczki cieczy. W miarę upływu czasu opis ten może

ulec zmianie. Dlatego właśnie w równaniu występują szybkości zmian opisywanych wielkości zarówno w czasie, jak i w przestrzeni.

Równanie Naviera–Stokesa ma szlachetny fizyczny rodowód. Opiera się na zastosowaniu dynamiki newtonowskiej do każdej pojedynczej cząsteczki cieczy (lub małego jej obszaru) i jest w tym kontekście wyrazem zasady zachowania pędu. Każda cząsteczka porusza się dlatego, że działają na nią siły, a z zasad dynamiki Newtona wynika, że jej przyspieszenie jest proporcjonalne do działającej siły. Głównymi siłami działającymi w tym wypadku są tarcie – wynikające z lepkości – i ciśnienie. Pojawiają się również siły będące efektem przyspieszenia cząsteczki. Samo równanie jest zgodne z tradycyjnym podejściem i traktuje ciecz jako nieskończenie podzielne kontinuum. W szczególności nie uwzględnia ono dyskretnej struktury atomowej cieczy w bardzo małych skalach.

Równania same w sobie mają niewielką wartość – trzeba jeszcze umieć je rozwiązać. W przypadku równania Naviera–Stokesa oznacza to konieczność wyznaczenia pola prędkości, czyli szybkości i kierunku każdej cząsteczki cieczy w każdym punkcie przestrzeni i w każdej chwili. Równanie wprowadza ograniczenia na te wielkości, ale nie podaje ich w sposób bezpośredni. Dopiero gdy go użyjemy, będziemy mogli powiązać przyszłe prędkości z obecnymi. Równania różniczkowe cząstkowe, takie jak równanie Naviera–Stokesa, mają wiele różnych rozwiązań – ba, nawet nieskończenie wiele. Trudno się temu dziwić, bo ciecze mogą przecież płynąć na wiele różnych sposobów. Przepływ po powierzchni karoserii samochodowej różni się od przepływu po skrzydłach samolotu. Z tego bogactwa możliwości musimy w jakiś sposób wybrać określone rozwiązanie i mamy dwie metody pozwalające tego dokonać – możemy sprecyzować warunki początkowe i warunki brzegowe.

Warunki początkowe określają pole prędkości w pewnej ustalonej chwili odniesienia; zwykle przyjmuje się, że jest to czas równy zeru. Z fizycznego punktu widzenia chodzi o to, że jeśli znamy pole prędkości w danej chwili, to równanie Naviera–Stokesa pozwoli w sposób jednoznaczny określić, jak wygląda to pole krótką chwilę później. Jeśli na początku popchniemy ciecz, to będzie się dalej poruszała zgodnie z prawami fizyki. W większości zastosowań

bardziej użyteczne są jednak warunki brzegowe, ponieważ określenie warunków początkowych dla rzeczywistej cieczy jest bardzo trudne, a poza tym nie do końca nadają się one do wykorzystania w wielu praktycznych zastosowaniach, na przykład takich jak projektowanie nadwozia samochodowego. Wówczas bowiem istotny jest kształt pojazdu. Lepka ciecz przyczepia się do powierzchni. Od strony matematycznej cechę tę modeluje się, określając prędkości na powierzchniach stanowiących brzeg obszaru zajętego przez ciecz, w którym obowiązuje nasze równanie. Możemy na przykład wprowadzić warunek, że na brzegu prędkość musi być równa zeru, lub zastosować inny wymóg, lepiej oddający rzeczywistość.

Nawet gdy określone są warunki początkowe i brzegowe, w wyjątkowych tylko wypadkach udaje się zapisać jawny wzór na pole prędkości, ponieważ równanie Naviera–Stokesa jest nieliniowe. Suma dwóch rozwiązań zwykle sama nie jest rozwiązaniem. To jeden z powodów, dla których problem trzech ciał omawiany w rozdziale 8 jest tak trudny – ale nie jest to jedyna przyczyna, ponieważ problem dwóch ciał jest również nieliniowy, a mimo to ma jawne rozwiązanie.

W wypadku zastosowań praktycznych możemy rozwiązać równanie Naviera–Stokesa na komputerze, zapisując pole prędkości w postaci ciągu liczb. Ciąg taki można przekształcić w elegancki wykres i wyliczyć na jego podstawie wielkości interesujące inżynierów, takie jak obciążenia działające na skrzydła samolotu. Ponieważ komputery nie potrafią przeprowadzać obliczeń na nieskończenie długich listach liczb, musimy zastąpić rzeczywisty przepływ jego dyskretnym przybliżeniem, czyli listą liczb będących wynikiem próbkowania przepływu w skończonej liczbie miejsc i chwil. Musimy przy tym koniecznie zrobić wszystko, co w naszej mocy, żeby takie przybliżenie było wystarczająco dobre.

Zwykle tworzy się w tym celu siatkę obliczeniową, dzieląc przestrzeń na olbrzymią liczbę małych obszarów, a następnie oblicza się prędkość tylko w punktach leżących w węzłach sieci. Taka siatka może się po prostu składać z kwadratów (lub sześcianów w trzech wymiarach), niczym szachownica, ale w przypadku samochodów i samolotów musi mieć bardziej skomplikowaną strukturę, z mniejszymi obszarami w pobliżu brzegu, by lepiej uchwycić

szczegóły przepływu. Można też zastosować siatkę dynamiczną, która zmienia kształt z upływem czasu. Ogólnie rzecz biorąc, przyjmuje się, iż czas płynie skokowo, z tym że długość każdego kroku może być taka sama lub zmieniać się w zależności od stanu obliczeń.

Podstawą działania większości metod numerycznych jest sposób zdefiniowania w rachunku różniczkowym pojęcia „szybkość zmiany". Przypuśćmy, że jakieś ciało przemieszcza się z jednego miejsca w drugie w bardzo krótkim czasie. W takim razie szybkość zmiany jego położenia – czyli prędkość – jest ilorazem zmiany położenia i czasu, jaki w tym czasie upłynął. Takie podejście obarczone jest niewielkim błędem, który maleje, gdy odcinek czasu staje się coraz krótszy. Możemy więc przybliżyć rzeczywistą szybkość zmiany, która jest jedną z wielkości wchodzących w skład równania Naviera–Stokesa, stosunkiem zmiany w przestrzeni do zmiany w czasie. W takiej postaci równanie mówi nam w zasadzie, jak należy przesunąć znany stan początkowy – określoną listę prędkości – o jeden krok czasowy w przyszłość. Następnie musimy powtórzyć takie obliczenia wielokrotnie, żeby przekonać się, co się wydarzy w bardziej odległej przyszłości. Istnieje podobny sposób na wyznaczanie przybliżonych rozwiązań, gdy szukana wielkość jest określona warunkami brzegowymi. Znamy również wiele zaawansowanych sposobów na obliczenie tego z większą dokładnością.

Jeśli zastosujemy gęstszą siatkę i krótsze odstępy czasu, to uzyskane przybliżenie będzie dokładniejsze, ale obliczenia będą trwały dłużej. Trzeba się więc zdecydować na jakiś kompromis między dokładnością a szybkością obliczeń. Na ogół przybliżone rozwiązanie wyznaczone komputerowo będzie zapewne wystarczająco dobre wtedy, gdy w przepływie nie występują jakieś istotne własności na obszarach mniejszych od rozmiaru elementów siatki. Wyróżniamy dwa główne rodzaje przepływów cieczy: laminarny i turbulentny. W przepływie laminarnym ruch przebiega gładko i poszczególne warstwy cieczy ślizgają się po sobie bez żadnych przeszkód. Tutaj więc powinna w zupełności wystarczyć niezbyt gęsta siatka. Przepływ turbulentny przebiega znacznie

gwałtowniej i ciecz miesza się wówczas w niezwykle skomplikowany sposób. W takich okolicznościach każda dyskretna siatka, nawet najgęstsza, bardzo łatwo może stać się źródłem problemów.

Jedną z cech charakterystycznych przepływu turbulentnego jest występowanie wirów, które mogą być naprawdę małe. Typowym przykładem takiego przepływu jest kaskada coraz mniejszych wirów. Większość szczegółów jest w tym wypadku mniejsza od rozmiaru elementów dowolnej siatki, którą można wykorzystać w praktyce. Aby obejść tę niedogodność, inżynierowie badający przepływ turbulentny posługują się często modelami statystycznymi. Innym źródłem zmartwień jest fakt, że fizyczny model ciągły może się okazać nieodpowiedni dla przepływu turbulentnego, ponieważ wiry mogą kurczyć się do rozmiarów atomowych. Z porównania wyników obliczeń numerycznych i doświadczeń fizycznych wynika, że równanie Naviera–Stokesa jest bardzo realistycznym i dokładnym modelem – tak dobrym, że w wypadku wielu zastosowań technicznych inżynierowie poprzestają na wykorzystaniu obliczeniowej mechaniki płynów, bo jej zastosowanie jest tańsze od kosztownych doświadczeń z modelami w tunelach aerodynamicznych. Jednak gdy stawką jest bezpieczeństwo ludzi, na przykład podczas projektowania samolotów, uzyskane wyniki wciąż sprawdza się doświadczalnie.

Równanie Naviera–Stokesa jest tak dokładne, że wydaje się, iż obowiązuje nawet w sytuacji, gdy zgodnie z regułami fizyki ma pełne prawo przestać działać, czyli w odniesieniu do przepływu turbulentnego. Tak jest przynajmniej wtedy, gdy daje się je rozwiązać wystarczająco dokładnie. Główny problem ma naturę praktyczną: gdy przepływ staje się turbulentny, metody numeryczne stosowane do rozwiązywania równania Naviera–Stokesa zabierają olbrzymią ilość czasu obliczeniowego. I zawsze pomijają jakieś małe struktury.

Matematycy czują się niepewnie, gdy najważniejsza informacja, jaką posiadają na temat jakiegoś problemu, opiera się na przybliżeniach. Milenijna nagroda związana z równaniem Naviera–Stokesa dotyka

jednej z kluczowych kwestii teoretycznych. Rozwiązanie tego równania byłoby potwierdzeniem naszego przeczucia, że metody numeryczne zazwyczaj działają bardzo dobrze. Istnieje pewna subtelna różnica między przybliżeniami stosowanymi w obliczeniach komputerowych, które skutkują wprowadzeniem niewielkich zmian do równań, a dokładnością wyniku, dotyczącą małych zmian w rozwiązaniu. Czy dokładna odpowiedź na przybliżone pytanie jest tym samym co przybliżona odpowiedź na dokładne pytanie? Czasami odpowiedź brzmi: „nie". Na przykład dokładny przepływ cieczy o bardzo małej lepkości nierzadko różni się od przybliżonego przepływu cieczy o zerowej lepkości.

Aby trochę lepiej zrozumieć te kwestie, wystarczy zauważyć coś tak prostego, że bez trudu można to przeoczyć – chodzi o spostrzeżenie, że dobrze jest mieć pewność, iż dokładne rozwiązanie w ogóle istnieje. Musi istnieć *coś*, czego przybliżeniem są obliczenia komputerowe. Właśnie ten fakt stał się motywacją do ustanowienia nagrody milenijnej związanej z równaniem Naviera–Stokesa. W oficjalnym opisie warunków przyznania tej nagrody, zamieszczonym na stronie internetowej Instytutu Claya, wymieniono cztery problemy. Aby zdobyć nagrodę, wystarczy rozwiązać jeden z nich. We wszystkich czterech punktach zakłada się, że ciecz jest nieściśliwa. Oto one:

1. *Istnienie i gładkość rozwiązań w trzech wymiarach*. W tym przypadku zakłada się, że ciecz wypełnia całą nieskończoną przestrzeń. Mając dane dowolne początkowe, gładkie pole prędkości, udowodnij, że dla wszystkich dodatnich czasów istnieje gładkie rozwiązanie równania odpowiadające podanemu polu początkowemu.
2. *Istnienie i gładkość rozwiązań w trójwymiarowym płaskim torusie*. To samo pytanie, ale teraz zakładamy, że przestrzeń jest płaskim torusem, czyli prostopadłościanem, w którym utożsamiono ze sobą przeciwległe ściany. Ta wersja pozwala ominąć potencjalne problemy, jakie mogą wynikać z istnienia nieskończonej dziedziny

w pierwszym punkcie, co tak naprawdę nie ma znaczenia i zupełnie niepotrzebnie może prowadzić do kłopotliwych sytuacji.

3. *Obalenie rozwiązań w trzech wymiarach*. Udowodnij, że twierdzenie z punktu pierwszego jest nieprawdziwe. Innymi słowy, znajdź takie pole początkowe, dla którego nie istnieje gładkie rozwiązanie dla wszystkich dodatnich czasów, i udowodnij to.

4. *Obalenie rozwiązań na trójwymiarowym płaskim torusie*. Udowodnij, że twierdzenie z punktu drugiego jest nieprawdziwe.

Te same problemy wciąż są nierozwiązane również dla równania Eulera, które jest takie samo jak równanie Naviera–Stokesa, z tą tylko różnicą, że nie uwzględnia lepkości. Jednak w wypadku równania Eulera nie wyznaczono żadnej nagrody za rozstrzygnięcie tych kwestii.

Główna trudność polega tu na tym, że rozważamy przepływ trójwymiarowy. Istnieje analogiczne równanie dla przepływu cieczy na płaszczyźnie. Z fizycznego punktu widzenia opisuje ono albo cienką warstwę cieczy między dwiema płaskimi płytami – oczywiście przy założeniu, że nie powodują one tarcia – albo wzorzec przepływu w trzech wymiarach, w którym ciecz porusza się tak samo wzdłuż całego układu płaszczyzn równoległych. W 1969 roku rosyjska matematyczka Olga Aleksandrowna Ładyżenskaja udowodniła dla dwuwymiarowego równania Naviera–Stokesa i dwuwymiarowego równania Eulera, że twierdzenia z punktów 1 i 2 są prawdziwe, a te z punktów 3 i 4 – niezgodne z prawdą.

Być może to dziwne, ale dowód okazał się trudniejszy dla równania Eulera, mimo że jest ono prostsze od równania Naviera–Stokesa, pomija bowiem wyrazy związane z lepkością. Przyczyna tego jest bardzo pouczająca. Lepkość „tłumi” kłopotliwe zachowania w rozwiązaniu, które ewentualnie mogłyby doprowadzić do powstania osobliwości uniemożliwiającej istnienie rozwiązania dla każdego czasu. Gdy pominiemy wyraz opisujący lepkość, tłumienie takie nie zachodzi, co przekłada się na problemy matematyczne w przeprowadzeniu dowodu istnienia rozwiązań.

Oprócz przedstawienia dowodu, że równanie Naviera–Stokesa ma rozwiązanie, Ładyżenskaja wniosła jeszcze inny ważny wkład w naszą wiedzę na ten temat: udowodniła, że pewne metody obliczeniowe mechaniki płynów pozwalają przybliżyć to równanie z dowolną dokładnością.

Przykłady opisane przez Instytut Claya odwołują się do przepływu cieczy nieściśliwej, ponieważ wiemy już, że w przypadku cieczy ściśliwej równania zachowują się w sposób trudny do opanowania. Na przykład równania wykorzystywane w procesie konstrukcji samolotów natrafiają na różnorodne problemy, gdy samolot porusza się z prędkością naddźwiękową. Chodzi tu o przekroczenie słynnej „bariery dźwięku", które sprawia duże kłopoty inżynierom projektującym naddźwiękowe odrzutowce. Problem ten ma związek ze ściśliwością powietrza. Gdy ciało porusza się w cieczy nieściśliwej, odpycha cząsteczki na boki, tak jak gdyby próbowało przedostać się przez pudło wypełnione kulkami łożyskowymi. Gdy dochodzi do nagromadzenia cząsteczek, spowalniają one ruch ciała. Jednak w cieczy ściśliwej, w której istnieje prędkość graniczna przemieszczania się fal – w naszym przykładzie jest to prędkość dźwięku – nic takiego się nie dzieje. Przy prędkościach naddźwiękowych powietrze nie jest odpychane na boki, ale gromadzi się z przodu samolotu, gdzie jego gęstość rośnie bez ograniczeń. W efekcie powstaje fala uderzeniowa. Z matematycznego punktu widzenia jest to nieciągłość w ciśnieniu powietrza, którego wartość w obszarze fali uderzeniowej nagle się zmienia. Na gruncie fizyki powiemy, że powstaje wówczas grom dźwiękowy – potężny huk. Jeśli powstawania fali uderzeniowej nie uwzględni się odpowiednio na etapie projektowania, może ona zniszczyć samolot, inżynierowie słusznie więc podchodzą do tego zagadnienia z dużą ostrożnością. Jednak prędkość dźwięku nie jest w istocie barierą, ale jedynie przeszkodą. Powstawanie fal uderzeniowych oznacza, że równania Naviera–Stokesa dla cieczy ściśliwych nie muszą mieć gładkich rozwiązań dla wszystkich wartości czasu, nawet w dwóch wymiarach.

Zatem znamy już odpowiedź na pytanie postawione przez ekspertów Instytutu Claya i jest to odpowiedź przecząca.

Matematyczny opis fal uderzeniowych jest rozległą dziedziną badań cząstkowych równań różniczkowych mimo faktu, że w tym przypadku rozwiązania się załamują. Chociaż samo równanie Naviera–Stokesa nie jest dobrym fizycznym modelem cieczy ściśliwych, model matematyczny można zmodyfikować, wprowadzając dodatkowe warunki, uwzględniające nieciągłości fal uderzeniowych. Jednak fale uderzeniowe nie powstają w cieczach nieściśliwych, a zatem przynajmniej w tym wypadku można mieć nadzieję, że rozwiązania będą istniały dla wszystkich chwil, bez względu na to, jak skomplikowany jest początkowy przepływ, byle tylko był gładki.

Dla trójwymiarowego równania Naviera–Stokesa znamy już nawet kilka rozwiązań. Jeśli wzorzec początkowego przepływu zawiera odpowiednio małe prędkości – a więc gdy przepływ ten jest bardzo leniwy – to spełnione są twierdzenia z punktów 1 i 2. Nawet gdy prędkości te są duże, to punkty 1 i 2 pozostają prawdziwe w pewnym niezerowym odcinku czasu. Być może nie istnieje rozwiązanie obowiązujące dla całej przyszłości, ale wiemy, że istnieje ono dla pewnego określonego okresu. To dziwne, bo mogłoby się przecież wydawać, że możemy powtórzyć cały proces, popychając rozwiązanie do przodu w czasie o niewielką wartość i wstawiając wynik końcowy jako nowy warunek początkowy. Problem z takim podejściem polega na tym, że przedziały czasu mogą kurczyć się tak szybko, iż pokonanie skończonego odcinka czasu może wymagać wykonania nieskończenie wielu kroków. Jeśli na przykład każdy kolejny krok trwa o połowę krócej od poprzedniego, a pierwszy zabiera, powiedzmy, 1 minutę, to cały proces skończy się po czasie $1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots$, wynoszącym w sumie 2 minuty. Jeśli rozwiązanie przestaje istnieć – w chwili obecnej jest to czysto hipotetyczne założenie, ale wciąż możemy je rozważać – to mówimy wówczas, że takie rozwiązanie *wybucha*. Czas, jaki upływa, zanim do tego dojdzie, nazywamy czasem wybuchu.

Zatem cztery pytania postawione na liście problemów milenijnych są w istocie pytaniami o to, czy rozwiązania mogą wybuchnąć. Jeśli

nie, to twierdzenia 1 i 2 są prawdziwe, a jeśli tak – to zgodne z prawdą są twierdzenia 3 i 4. A może rozwiązania mogą wybuchnąć w dziedzinie nieskończonej, ale już nie w skończonej? Przy okazji zauważmy, że jeśli odpowiedź na pytanie 1 brzmi: „tak", to odpowiedź na pytanie 2 również jest twierdząca, ponieważ dowolny wzorzec przepływu w płaskim torusie możemy zinterpretować jako okresowy przestrzennie wzór przepływu w całej nieskończonej przestrzeni. Można sobie wyobrazić, że wypełniamy całą przestrzeń kopiami rozważanego równoległoboku i w każdym z nich powielamy taki sam wzór przepływu. Reguły sklejania zdefiniowane dla torusa sprawią, że przepływ pozostanie gładki przy przejściu przez płaskie powierzchnie graniczne. Podobnie jeśli odpowiedź na pytanie 4 brzmi: „tak", to z tego samego powodu odpowiedź na pytanie 3 również jest twierdząca. Musimy jedynie sprawić, żeby stan początkowy był okresowy. O ile nam jednak obecnie wiadomo, odpowiedź na pytanie 2 może być twierdząca, natomiast odpowiedź na pytanie 1 może brzmieć: „nie".

Wiemy jednak coś bardzo interesującego na temat wybuchania rozwiązań. Jeśli istnieje jakieś rozwiązanie o skończonym czasie wybuchu, to maksymalna prędkość cieczy we wszystkich punktach przestrzeni musi być dowolnie duża. Mogłoby do tego na przykład dojść, gdyby utworzyła się struga, której prędkość zwiększałaby się tak szybko, że po skończonym czasie wzrosłaby do nieskończoności.

Nie są to czysto hipotetyczne zastrzeżenia. Znane są przypadki takiego osobliwego zachowania innych równań klasycznej fizyki matematycznej. Niezwykły tego przykład możemy znaleźć w mechanice nieba. W 1988 roku Zhihong Xia udowodnił, że w pewnej konfiguracji początkowej pięciu mas punktowych w trójwymiarowej przestrzeni, zachowujących się zgodnie z prawami newtonowskiej grawitacji, cztery cząstki uciekają do nieskończoności po upływie skończonego czasu – jest to pewna forma wybuchu – a piąta podlega coraz gwałtowniejszym drganiom. Wcześniej Joseph Gerver zwrócił uwagę na to, że pięć ciał na płaszczyźnie może uciec do nieskończoności w skończonym czasie, ale nie udało mu się

przedstawić pełnego dowodu takiego scenariusza. W 1989 roku udowodnił, że na płaszczyźnie może dojść do tego rodzaju ucieczki, gdy liczba ciał jest odpowiednio duża.

To niezwykłe, że takie zachowanie jest możliwe, bo w tego typu układach obowiązuje zasada zachowania energii. Jeżeli wszystkie ciała poruszają się z dowolnie dużą prędkością, to musi przecież dojść do wzrostu energii kinetycznej. Aby to wyjaśnić, należy zauważyć, że dochodzi wówczas również do zmniejszenia energii potencjalnej, a w przypadku cząstki punktowej całkowita grawitacyjna energia potencjalna jest nieskończona. Ciała muszą również przestrzegać zasady zachowania momentu pędu, ale by ją spełnić, wystarczy, że niektóre z nich będą się poruszały coraz szybciej po coraz mniejszych okręgach.

Na gruncie fizyki wiąże się z tym słynny manewr asysty grawitacyjnej, z którego korzysta się obecnie standardowo podczas wysyłania sond kosmicznych do odległych obiektów Układu Słonecznego. Dobrym przykładem może być wystrzelona przez NASA sonda Galileo, która miała dotrzeć do Jowisza, by zbadać tego gazowego olbrzyma oraz jego liczne satelity. Wystrzelono ją w 1989 roku, ale do celu dotarła dopiero w 1995 roku. Jednym z powodów, dla których podróż trwała tak długo, było to, że sonda wcale nie leciała po najkrótszej drodze. Chociaż orbita Jowisza leży na zewnątrz orbity ziemskiej, na samym początku sondę skierowano do środka Układu Słonecznego, w stronę Wenus. Po przejściu w pobliżu Wenus sonda ponownie minęła Ziemię i poleciała dalej, żeby przyjrzeć się planetoidzie 951 Gaspra. Potem zawróciła ponownie w kierunku Ziemi, jeszcze raz okrążyła naszą planetę i w końcu wyruszyła w stronę Jowisza. Po drodze zbliżyła się do jeszcze jednej planetoidy noszącej nazwę 243 Ida i odkryła, że ma ona własny maleńki księżyc, który otrzymał nazwę Daktyl.

Dlaczego wybrano tak zawiłą trajektorię? Dlatego, że przy każdym minięciu ciała niebieskiego sonda Galileo zwiększała swoją energię, a więc i prędkość. Wyobraźmy sobie sondę, która leci w kierunku zbliżającej się do niej planety – nie kursem kolizyjnym, ale bardzo blisko jej powierzchni – zakręca za planetą i zostaje wyrzucona w przestrzeń kosmiczną. Gdy sonda przelatuje za planetą, oba ciała

przyciągają się grawitacyjnie. Tak naprawdę przyciągają się przez cały czas, ale na tym odcinku przyciąganie grawitacyjne jest najsilniejsze, wywiera więc największy efekt. Grawitacja planety zwiększa prędkość sondy. Ponieważ energia musi być zachowana, sonda zmniejsza w nieznacznym stopniu prędkość ruchu orbitalnego planety. Sonda ma jednak bardzo małą masę w porównaniu z olbrzymią masą planety, zatem jej wpływ na ruch planety jest zaniedbywalnie mały. Nie można tego natomiast powiedzieć o wpływie planety na sondę – sonda zdecydowanie zwiększa prędkość.

Sonda Galileo zbliżyła się do powierzchni Wenus na odległość 16 tysięcy kilometrów, w wyniku czego zwiększyła swoją prędkość o 2,23 kilometra na sekundę. Następnie minęła Ziemię w odległości 960 kilometrów, a potem jeszcze raz w odległości 300 kilometrów, zwiększając swoją prędkość o kolejne 3,7 kilometra na sekundę. Takie manewry były niezbędne, by mogła dotrzeć do Jowisza, ponieważ korzystając wyłącznie ze swoich silników rakietowych, nie byłaby w stanie dolecieć tam najkrótszą drogą. Pierwotnie planowano, że w sondzie zostaną zamontowane silniki na ciekły wodór *Centaur-G* i poleci ona prosto w kierunku Jowisza. Jednak po katastrofie promu kosmicznego *Challenger* tuż po starcie zakazano dalszego stosowania silników *Centaur-G* i trzeba było zmienić plany. Sonda musiała dolecieć do celu na znacznie słabszym silniku na paliwo stałe. Jej misja była olbrzymim sukcesem, a dodatkową korzyścią naukową z jej przeprowadzenia była możliwość obserwacji uderzenia komety Shoemaker-Levy 9 w Jowisza w 1994 roku, w czasie gdy sonda zmierzała dopiero w jego kierunku.

W swoim scenariuszu Zhihong Xia również wykorzystuje manewr asysty grawitacyjnej. Cztery planety o równej masie tworzą w nim dwie bliskie pary orbitujące wokół wspólnego środka masy, z tym że orbity obu par leżą na dwóch płaszczyznach równoległych[80]. Takie dwuciałowe rakiety grają w kosmicznego tenisa piątym, lżejszym ciałem, które odbija się pod kątem prostym tam i z powrotem między obiema płaszczyznami. Konfiguracja całego układu jest taka, że za każdym razem gdy „piłka tenisowa” mija parę planet, efekt

asysty grawitacyjnej zwiększa jej prędkość i odpycha parę planet na zewnątrz wzdłuż linii łączącej obie pary, co sprawia, że kort tenisowy bezustannie się wydłuża i gracze znajdują się w coraz większej odległości od siebie. Energia i pęd pozostają cały czas w równowadze, ponieważ pary planet zbliżają się coraz bardziej do siebie i coraz szybciej krążą wokół wspólnego środka masy. Przy odpowiedniej konfiguracji początkowej obie pary planet odsuwają się od siebie coraz szybciej, a ich prędkość wzrasta w tak dużym tempie, że po upływie skończonego czasu oddalają się od siebie na nieskończoną odległość. Tymczasem piłka tenisowa kursuje między nimi coraz szybciej. W swojej wersji scenariusza ucieczki Gerver również wykorzystuje efekt asysty grawitacyjnej.

Czy takie sztuczki ze znikaniem mają w ogóle jakieś znaczenie w odniesieniu do prawdziwych ciał niebieskich? Nie, jeśli rozumieć je dosłownie. Aby mogły zadziałać, ciała muszą być masami punktowymi. Gdy rozważa się różne problemy z zakresu mechaniki nieba, założenie takie bardzo często jest sensowne, ale nie można go stosować, gdy ciała zbliżają się do siebie na dowolnie małą odległość. Gdyby doszło do czegoś takiego w przypadku ciał o skończonym rozmiarze, po prostu by się w końcu zderzyły. Poza tym efekty relatywistyczne nie pozwoliłyby na rozwinięcie prędkości większej od prędkości światła, wprowadzając odpowiednie poprawki do praw grawitacji. W dodatku wymagana początkowa konfiguracja i przyjęte założenie, że masy niektórych ciał są jednakowe, i tak oznacza, że wystąpienie takiego układu w rzeczywistości byłoby niezwykle mało prawdopodobne. Niemniej rozważając takie dziwne przykłady, możemy się przekonać, że chociaż równania mechaniki nieba w większości wypadków bardzo dobrze opisują rzeczywistość, to jednak mogą się w nich pojawiać skomplikowane osobliwości, które sprawiają, że rozwiązania nie muszą wcale istnieć dla wszystkich czasów. Niedawno uświadomiono sobie również, że efekt asysty grawitacyjnej w układach potrójnych gwiazd – czyli w takich, w których trzy gwiazdy krążą wokół siebie po skomplikowanych orbitach – może doprowadzić do wyrzucenia jednej z nich w kosmos z dużą prędkością. Niewykluczone więc, że Galaktykę – a nawet przestrzeń międzygalaktyczną – przemierzają niezliczone samotne

gwiazdy wyrzucone ze swoich układów przez siostry i błądzą tak w zimnie, samotne, niechciane i przez nikogo niezauważane.

Gdy równanie różniczkowe zachowuje się tak dziwnie, że jego rozwiązania stają się bezsensowne po upływie pewnego skończonego czasu, mówimy, iż występuje w nich osobliwość. Wspomniane przed chwilą prace na temat układów wielociałowych dotyczą tak naprawdę różnych rodzajów osobliwości. Pytania z listy problemów milenijnych na temat równania Naviera–Stokesa odnoszą się do możliwości powstania osobliwości w wyniku przyjęcia określonych warunków początkowych dla cieczy wypełniającej albo całą przestrzeń, albo płaski torus. Jeśli w skończonym czasie pojawi się osobliwość, to wynik zapewne wybuchnie – można by tego uniknąć, gdyby osobliwość w jakiś sposób sama się rozwiązała, ale wydaje się to mało prawdopodobne.

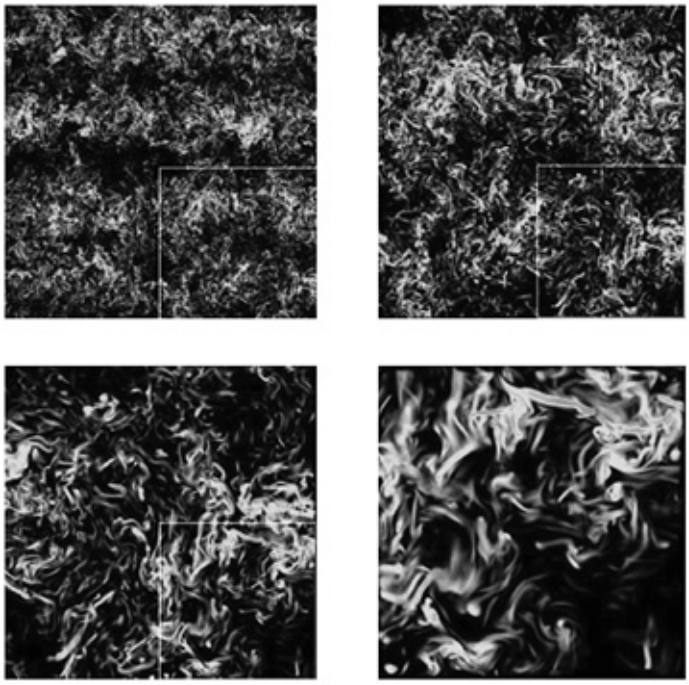
Istnieją dwie główne metody zmierzenia się z tak postawionymi pytaniami. Możemy spróbować udowodnić, że osobliwości nigdy nie powstają, albo możemy wyszukiwać kolejne odpowiednie warunki początkowe. Metody numeryczne będą pomocne w obu wypadkach, gdyż pozwalają zrozumieć użyteczne ogólne cechy przepływów i dostarczają ważnych wskazówek na temat rodzaju potencjalnych osobliwości. Ponieważ jednak w obliczeniach numerycznych zawsze istnieje pewne ryzyko wystąpienia niedokładności, wskazówki takie należy traktować z ostrożnością i znaleźć dla nich ściślejsze uzasadnienie.

Matematycy próbujący udowodnić regularność – czyli brak osobliwości – stosują różnorodne metody pozwalające sprawować kontrolę nad przepływem. Są to między innymi skomplikowane metody szacowania, jak duże lub małe wartości mogą osiągnąć pewne kluczowe zmienne, i dużo bardziej abstrakcyjne techniki. Często stosowane podejście wykorzystuje tak zwane rozwiązania słabe, które nie są tak naprawdę wcale przepływami, ale bardziej ogólnymi strukturami matematycznymi przejawiającymi pewne własności przepływów. Wiadomo na przykład, iż zbiór osobliwości

słabego rozwiązania trójwymiarowych równań Naviera–Stokesa jest zawsze mały, z tym że określenie to ma pewne ścisłe znaczenie.

Zbadano wiele różnych scenariuszy mogących prowadzić do powstania osobliwości. Autorem standardowego modelu turbulencji jako kaskady coraz mniejszych wirów jest Andriej Kołmogorow, który zaproponował go w 1941 roku. Wysunął wówczas hipotezę, że w bardzo małych skalach wszystkie rodzaje turbulencji wyglądają bardzo podobnie. Przyjął na przykład, że proporcje wirów danego rozmiaru wynikają z jakiegoś powszechnie obowiązującego prawa. Obecnie wiemy, że gdy wiry stają się coraz mniejsze, zmieniają swój kształt – wydłużają się i robią się coraz cieńsze, tworząc włókna. Z zasady zachowania momentu pędu wynika, że wirowość – czyli to, jak bardzo wir jest skręcony – musi rosnąć. Zjawisko to, nazywane rozciąganiem wiru, może prowadzić do powstania osobliwości – na przykład gdyby bardzo małe wiry mogły się stawać nieskończenie długie po upływie skończonego czasu, a wirowość osiągałaby w niektórych miejscach wartość nieskończoną.

Na rycinie 43 pokazano bardzo duże powiększenie symulacji przepływu turbulentnego, którą przeprowadził Pablo Mininni wraz z kolegami, wykorzystując do tego program VAPOR (ang. „para”; skrót od ang. *Visualization and Analysis Platform for Ocean, Atmosphere and Solar Research* – platforma do wizualizacji i analizy wyników badań oceanu, atmosfery i Słońca). Na zamieszczonych ilustracjach pokazano natężenie wirowości – czyli tego, jak szybko wiruje ciecz. Widać na nich, jak powstają włókna wirowe – to te długie i cienkie struktury – i jak mogą się grupować, tworząc wzory w większych skalach. Program używany przez zespół Mininniego umożliwia przeprowadzenie symulacji w sieci sześciennej obejmującej ponad 3 miliardy punktów.



Ryc. 43. Kolejne powiększenia symulacji przepływu turbulentnego przeprowadzonej z wykorzystaniem programu komputerowego VAPOR

W artykule poświęconym temu problemowi, który został zamieszczony na stronie Instytutu Claya, Charles Fefferman pisze[81]:

> Istnieje wiele fascynujących problemów i hipotez związanych z zachowaniem się rozwiązań równań Eulera i Naviera–Stokesa. […] Ponieważ nie wiemy nawet, czy rozwiązania te istnieją, nasza wiedza jest na bardzo prymitywnym poziomie. Standardowe metody rozwiązywania równań [różniczkowych cząstkowych] wydają się niewystarczające do rozstrzygnięcia tej kwestii. Potrzebne są zapewne jakieś głębokie, nowe idee.

Złożoność przepływu widoczna na ilustracjach takich jak te zamieszczone na rycinie 43 pozwala sobie uświadomić, jak duże trudności możemy napotkać, szukając takich nowych idei.

Matematycy jednak się tym nie zrażają i mężnie maszerują dalej, poszukując w tej złożoności prostych zasad.

80 Zhihong Xia, *The existence of noncollision singularities in Newtonian systems* (*Istnienie bezkolizyjnych osobliwości w układach newtonowskich*), „Annals of Mathematics" 1992, tom 135, s. 411–468.

81 Zob.: http://www.claymath.org/millennium/Navier-Stokes_Equations/.

13. Kwantowa zagadka

Hipoteza luki masowej

Kilka kilometrów na północ od Genewy granica między Szwajcarią i Francją gwałtownie skręca. Na powierzchni widać w tym miejscu tylko wiejskie drogi i niewielkie miejscowości, ale pod ziemią, na głębokości od 50 do 175 metrów, znajduje się największy przyrząd naukowy na świecie. Tworzy go gigantyczny okrągły tunel o średnicy ponad 8 kilometrów, połączony z drugim okrągłym tunelem około czterech razy mniejszym. Większa część tunelu znajduje się na terytorium Francji, ale dwa jego odcinki przebiegają pod terenem należącym do Szwajcarii. Środkiem tunelu biegną dwie rury przecinające się w czterech miejscach.

Jest to Wielki Zderzacz Hadronów (w skrócie LHC, od ang. *Large Hadron Collider*), którego budowa kosztowała 7,5 miliarda euro. Za tę kwotę uczeni skonstruowali urządzenie umożliwiające rozszerzanie granic fizyki cząstek. Najważniejszym celem pracujących tu 10 tysięcy naukowców ze 100 krajów było znalezienie bozonu Higgsa – albo nieznalezienie go, gdyby się okazało, że natura jednak go nie potrzebuje. Cząstki tej brakowało im do skompletowania pełnego Modelu Standardowego fizyki cząstek, w którego wynika, że wszystko we Wszechświecie zbudowane jest z 17 różnych cząstek elementarnych. Według teorii wysuniętej przez uczonych, bozon Higgsa jest odpowiedzialny za nadanie wszystkim cząstkom masy.

W grudniu 2011 roku zespoły detektorów ATLAS i CMS, prowadzące doświadczenia w Wielkim Zderzaczu Hadronów, znalazły niezależnie dowody wskazujące na istnienie bozonu Higgsa o masie około 125 GeV (gigaelektronowoltów – w fizyce cząstek jednostki tej

używa się zamiennie na określanie ilości masy i energii, ponieważ, jak wiadomo, wielkości te są sobie równoważne). Czwartego lipca 2012 roku przedstawiciele CERN-u (europejskiego laboratorium, w ramach którego działa LHC) oznajmili uczonym i dziennikarzom tłumnie wypełniającym aulę ośrodka, że natura stanęła po stronie cząstki Higgsa. Oba zespoły zgromadziły olbrzymią ilość dodatkowych danych i ryzyko, że zaobserwowano jakąś przypadkową fluktuację, a nie nową cząstkę przypominającą cząstkę Higgsa, spadło do poziomu prawdopodobieństwa mniejszego niż 1 na 2 miliony. Zgodnie z przyjętymi zasadami w fizyce cząstek wymaga się osiągnięcia właśnie takiego poziomu ufności, zanim można będzie wznosić toasty.

Uczeni będą teraz musieli przeprowadzić dalsze doświadczenia, żeby upewnić się, iż nowa cząstka ma wszystkie własności, jakie powinny cechować teoretyczny bozon Higgsa. Teoria przewiduje na przykład, że cząstka Higgsa musi mieć zerowy spin. W chwili ogłoszenia wyników doświadczeń z badań wynikało, że spin nowo odkrytej cząstki może mieć wartość 0 lub 2. Istnieje także możliwość, że bozon Higgsa może się w istocie okazać cząstką złożoną, zbudowaną z mniejszych cząstek, lub że uczeni odkryli jedynie pierwszego przedstawiciela całej nowej rodziny cząstek Higgsa. Możemy się zatem spodziewać, że to odkrycie potwierdzi poprawność obecnego modelu cząstek elementarnych albo dostarczy nowych informacji, które ostatecznie doprowadzą do opracowania lepszej teorii.

Ostatni z siedmiu problemów milenijnych jest ściśle związany z Modelem Standardowym i bozonem Higgsa. Dotyczy najważniejszego otwartego pytania kwantowej teorii pola, która jest formalizmem matematycznym używanym w fizyce cząstek. Chodzi o tak zwaną hipotezę luki masowej, z której wynika minimalna wartość masy, jaką może mieć cząstka elementarna. Jest to tylko jeden przykładowy problem wybrany z całej grupy wielkich nierozstrzygniętych kwestii dotyczących tego głębokiego i zupełnie nowego działu fizyki matematycznej. Ma on związek z różnorodnymi obszarami badań naukowych, od granic matematyki teoretycznej po

prace mające na celu unifikację dwóch głównych teorii fizycznych: ogólnej teorii względności i kwantowej teorii pola.

W klasycznej mechanice newtonowskiej podstawowymi wielkościami fizycznymi są przestrzeń, czas i masa. Zakłada się w niej, że przestrzeń jest trójwymiarową przestrzenią euklidesową, czas – jednowymiarową wielkością niezależną od przestrzeni, a masa oznacza obecność materii. Pod wpływem działania sił masy zmieniają swoje położenie w przestrzeni, a szybkość zachodzenia tych zmian mierzy się względem czasu. Zasady dynamiki Newtona opisują, jak przyspieszenie ciała (szybkość zmiany prędkości, która sama jest szybkością zmiany położenia) zależy od jego masy i działającej siły.

Najwspanialszą klasyczną teorią przestrzeni, czasu i materii są równania elektromagnetyzmu Jamesa Clerka Maxwella[82], tworzące elegancki układ równań unifikujący dwie siły natury, które wcześniej uważano za zupełnie różne. Zamiast elektryczności i magnetyzmu pojawiło się jedno pole elektromagnetyczne. Pole to przenika całą przestrzeń, jak gdyby Wszechświat był wypełniony jakąś niewidzialną cieczą. W każdym punkcie przestrzeni możemy zmierzyć siłę i kierunek tego pola, jak gdyby ta wyimaginowana ciecz płynęła zgodnie z matematycznymi prawidłowościami. Czasami możemy rozdzielić pole elektromagnetyczne na jego dwie części składowe – pole elektryczne i magnetyczne – ale poruszające się pole magnetyczne wytwarza pole elektryczne, i odwrotnie, a zatem w układach dynamicznych oba pola należy połączyć w jedno, bardziej złożone.

Na początku XX stulecia ten pokrzepiający obraz świata fizycznego, w którym podstawowe pojęcia naukowe przypominają to, co postrzegamy naszymi zmysłami, uległ radykalnym zmianom. Fizycy zdali sobie wtedy sprawę, że w bardzo małych skalach, tak małych, że nie można ich było badać żadnymi dostępnymi wówczas mikroskopami, materia wcale nie jest taka, jak sobie wszyscy wyobrażali. Chemicy i fizycy zaczęli poważnie rozważać dość szaloną teorię, która swymi korzeniami sięgała zamierzchłej przeszłości sprzed dwóch tysięcy lat – filozoficznych rozważań Demokryta

w starożytnej Grecji oraz innych uczonych w Indiach. Chodziło o pogląd głoszący, że chociaż wydaje się, iż świat tworzą niezliczone, różnorodne substancje, to w istocie cała materia zbudowana jest z maleńkich cząstek – atomów. Określenie to pochodzi od greckiego słowa „niepodzielny".

W XIX stuleciu chemicy odkryli dowody pośrednie potwierdzające istnienie atomów – fakt, że pierwiastki łączą się ze sobą w bardziej złożone cząsteczki w ściśle określonych proporcjach, bardzo często bliskich liczb całkowitych. John Dalton wysnuł z tych obserwacji prawo stosunków wielokrotnych i zasugerował, że jego wyjaśnieniem mogą być atomy. Gdyby każdy związek chemiczny składał się ze stałej liczby różnych atomów, to takie prawo wynikałoby z tego faktu w sposób zupełnie naturalny. Obecnie wiemy na przykład, że każda cząsteczka dwutlenku węgla składa się z dwóch atomów tlenu i jednego atomu węgla, a więc ich stosunek będzie zawsze równy dwa do jednego. Pojawiają się jednak pewne komplikacje: różne atomy mają różne masy, a wiele pierwiastków występuje w postaci cząsteczek utworzonych z kilku atomów – na przykład cząsteczka tlenu składa się z dwóch atomów tlenu. Ktoś, kto tego nie wie, może sądzić, że masa atomu tlenu jest dwukrotnie większa niż w rzeczywistości. W dodatku niektóre pierwiastki są w istocie mieszaniną różnych izotopów – czyli różnych struktur atomowych. Chlor na przykład występuje w naturze jako mieszanka dwóch stabilnych form, które obecnie nazywamy chlorem-35 i chlorem-37, w proporcjach wynoszących odpowiednio około 76 i 24 procent. Zatem zmierzoną doświadczalnie „masę atomową" chloru, wynoszącą 35,45, na wczesnych etapach rozwoju teorii atomowej traktowano jako dowód na to, że „atom chloru składa się z trzydziestu pięciu i pół atomu wodoru". Oznaczałoby to, że atom wcale nie jest niepodzielny. Na początku XX wieku większość uczonych wciąż uważała, że uznanie teorii atomowej za poprawną było zbyt daleko idącym posunięciem i zebrane dowody liczbowe są zbyt słabe, by mogły uzasadniać jej przyjęcie.

Niektórzy naukowcy, szczególnie Maxwell i Ludwig Boltzmann, wyprzedzali swoich współczesnych i nie mieli wątpliwości, że gazy są rozproszonymi zbiorami cząsteczek, które powstają przez połączenie

atomów. Wydaje się, że ich kolegów przekonało ostatecznie przedstawione przez Alberta Einsteina wyjaśnienie ruchów Browna, które są widocznymi pod mikroskopem przypadkowymi ruchami maleńkich cząsteczek unoszących się w cieczy. Einstein doszedł do wniosku, że ruchy te muszą być wynikiem zderzeń z przypadkowo poruszającymi się cząsteczkami cieczy, i przeprowadził obliczenia ilościowe potwierdzające ten pogląd. Jean Perrin potwierdził te przewidywania doświadczalnie w 1908 roku. Możliwość zobaczenia efektów działania tych rzekomo niepodzielnych cząstek materii i wysunięcie ilościowych przewidywań okazało się dużo bardziej przekonujące od filozoficznych rozważań i analizy dziwnych związków liczbowych. W obliczu takich argumentów uczeni przyjęli w końcu do wiadomości istnienie atomów.

Gdy świat naukowy oswajał się dopiero z koncepcją atomów, niektórzy uczeni zaczęli już sobie zdawać sprawę z tego, że atomy wcale nie są niepodzielne, że mają jakąś strukturę i można wybijać z nich małe kawałki materii. W 1897 roku Joseph John Thomson przeprowadzał doświadczenia z tak zwanym promieniowaniem katodowym i odkrył, że atomy można zmusić do emisji jeszcze mniejszych cząstek – elektronów. W istocie stwierdził nawet coś więcej, a mianowicie to, że atomy różnych pierwiastków emitują te same cząstki. Wykorzystując pole magnetyczne, Thomson pokazał, że elektrony mają ujemny ładunek elektryczny. Ponieważ atomy są elektrycznie obojętne, oznaczało to, że musi również istnieć jakaś inna część atomu obdarzona ładunkiem dodatnim. W ten sposób Thomson doszedł do swojego modelu atomu, który bywa nazywany modelem „ciasta z rodzynkami”, ponieważ zakłada on, że atom przypomina dodatnio naładowane ciasto wypełnione rodzynkami o ładunku ujemnym. Jednak w 1909 roku Ernest Rutherford, jeden z jego byłych studentów, przeprowadził doświadczenia, które pokazały, że większa część masy atomu skupiona jest w pobliżu środka. Taki obraz w niczym nie przypomina ciasta.

Czy tak małe obszary przestrzeni można badać doświadczalnie? Wyobraźmy sobie pewien teren, na którym mogą stać jakieś budynki

i inne konstrukcje, ale wcale nie muszą. Nie możemy wejść na ten teren, a wokół panuje całkowita ciemność, nic więc nie widzimy. Mamy jednak ze sobą strzelbę i dużo pudełek z nabojami. Możemy strzelać na chybił trafił w kierunku interesującego nas obszaru i obserwować, w jakim kierunku wylatują kule. Jeżeli nasz teren przypomina ciasto z rodzynkami, większość kul przeleci przez niego na wylot. Jeśli zaś od czasu do czasu będziemy musieli zrobić unik przed kulą, która odbiła się rykoszetem i pędzi wprost na nas, będzie to oznaczało, że trafiliśmy w coś twardego. Możemy oszacować rozmiar takiego twardego obiektu, badając, jak często kule wylatują pod określonym kątem.

Kulami Rutherforda były cząstki alfa, czyli jądra atomów helu, a terenem, który ostrzeliwał – arkusz cienkiej złotej folii. Dzięki pracom Thomsona uczeni wiedzieli, że elektronowe rodzynki mają bardzo małą masę, a to oznaczało, że prawie cała masa atomu powinna być zawarta w cieście. Gdyby to ciasto nie zawierało żadnych grudek, większość cząstek alfa powinna przelecieć przez nie na wylot – tor ich ruchu bardzo rzadko ulegałby odchyleniu, a jeśli już, to o niezbyt duży kąt. Okazało się jednak, że niewielki, ale znaczący ułamek wystrzelonych cząstek ulega dużym odchyleniom. Zatem porównanie do ciasta z rodzynkami nie może oddawać rzeczywistości. Rutherford zaproponował inny model, który obecnie wciąż bywa często używany, chociaż uczeni przedstawili już nowocześniejsze obrazy atomu – chodzi o tak zwany model planetarny. Przyjmuje się w nim, że atom przypomina Układ Słoneczny – w jego środku znajduje się duże jądro, czyli jego „słońce”, wokół którego, na podobieństwo planet, krążą elektrony. Zatem tak jak Układ Słoneczny, wnętrze atomu składa się głównie z pustej przestrzeni.

Rutherford kontynuował badania i znalazł dowody na to, że jądro składa się z dwóch różnych rodzajów cząstek: protonów, obdarzonych dodatnim ładunkiem elektrycznym, i neutronów, które są elektrycznie obojętne. Obie te cząstki mają podobną masę, około 1800 razy większą od masy elektronu. Atomy nie są więc wcale niepodzielne. Składają się z jeszcze mniejszych cząstek subatomowych. Teoria ta wyjaśnia, skąd w opisie pierwiastków

chemicznych pojawiają się charakterystyczne liczby całkowite – wynikają one z liczby protonów i neutronów. Tłumaczy również istnienie izotopów – jeśli dodamy lub odejmiemy kilka neutronów, to zmienimy masę atomu, ale ładunek jądra pozostanie taki sam, a zatem nie zmieni się również liczba elektronów, która jest równa liczbie protonów. Za własności chemiczne atomu odpowiedzialne są głównie jego elektrony. Na przykład chlor-35 ma 17 protonów, 17 elektronów i 18 neutronów, natomiast chlor-37 ma 17 protonów, 17 elektronów i 20 neutronów. Wartość 35,45 występuje więc dlatego, że chlor w stanie naturalnym jest mieszaniną tych dwóch izotopów.

Na początku XX wieku pojawiła się nowa teoria opisująca materię w skalach cząstek subatomowych. Chodzi o mechanikę kwantową, która całkowicie odmieniła fizykę. Teoria ta przewidywała wiele nowych zjawisk i w krótkim czasie wiele z nich udało się zaobserwować w laboratoriach. Wyjaśniła dużo dziwnych i zagadkowych obserwacji. Przewidziała istnienie nowych cząstek elementarnych. Poza tym wynikało z niej, że klasyczny obraz naszego Wszechświata, który dotychczas tak świetnie zgadzał się z obserwacjami, jest jednak błędny. To, co dostrzegamy w naszej ludzkiej skali, jest kiepskim modelem rzeczywistości na jej najbardziej podstawowym poziomie.

W fizyce klasycznej materia składa się z cząstek, a światło jest falą. W mechanice kwantowej również światło składa się z cząstek – fotonów – natomiast cząstki materii, na przykład elektrony, mogą się czasem zachowywać jak fale. Obowiązujące dotychczas wyraźne rozróżnienie między falami i cząstkami nie tyle uległo rozmyciu, ile całkowicie znikło, zastąpione pojęciem dualizmu korpuskularno-falowego. Planetarny model atomu, rozumiany dosłownie, nie sprawdził się zbyt dobrze, zaproponowano więc nowy obraz. W nowym ujęciu elektrony nie krążą wokół jądra niczym planety, ale tworzą wokół niego rozmyty obłok. Nie jest to jednak chmura jakiejś materii, ale prawdopodobieństwa. Jej gęstość odpowiada prawdopodobieństwu znalezienia elektronu w danym obszarze.

Oprócz protonów, neutronów i elektronów fizycy znali jeszcze tylko jeden rodzaj cząstek subatomowych – fotony. Wkrótce jednak pojawiły się następne. Zaobserwowano przypadki, w których

wydawało się, że zasada zachowania energii nie jest spełniona. Wolfgang Pauli zaproponował ich wyjaśnienie, postulując istnienie neutrina – niewidocznej i praktycznie niemożliwej do wykrycia cząstki, która unosiłaby ze sobą brakującą energię. Okazało się, iż jest ona na tyle wykrywalna, że jej istnienie udało się potwierdzić w 1956 roku. I wtedy ruszyła lawina. Niedługo potem pojawiły się piony, miony i kaony – te ostatnie odkryto w wyniku obserwacji promieniowania kosmicznego. Tak narodziła się fizyka cząstek. Zajmujący się nią uczeni w dalszym ciągu badali niewiarygodnie małe skale przestrzenne taką samą metodą jak Rutherford: aby odkryć, co znajduje się w środku tych obszarów, ciskali w nie dużą liczbą cząstek i obserwowali, jak się odbijają. Wybudowano i uruchomiono specjalne duże akceleratory cząstek, które w zasadzie są działami wystrzeliwującymi kule. Powstał liniowy akcelerator imienia Stanforda o długości trzech kilometrów. Aby uniknąć konieczności konstruowania akceleratorów o długościach przekraczających rozmiary kontynentów, uczeni wygięli je w okrąg, dzięki czemu cząstki mogą krążyć w ich wnętrzu, rozpędzając się do olbrzymich prędkości. Wybudowanie takich akceleratorów było bardziej skomplikowane od strony technicznej, ponieważ cząstki krążące po okręgu wypromieniowują energię, ale udało się pokonać związane z tym problemy.

Pierwszym owocem tych prac był bezustannie powiększający się katalog cząstek wyglądających na cząstki elementarne. Enrico Fermi dał wyraz swojej frustracji tym faktem, stwierdzając: „Gdybym potrafił spamiętać nazwy wszystkich tych cząstek, zostałbym botanikiem”. Jednak od czasu do czasu pojawiały się nowe idee z zakresu teorii kwantowej, za sprawą których lista ta ulegała skróceniu – działo się tak za każdym razem, gdy fizycy wysuwali hipotezę istnienia coraz mniejszych cząstek unifikujących zaobserwowane już wcześniej struktury.

Na początku mechanika kwantowa opisywała pojedyncze obiekty przypominające cząstkę bądź falę. Nikomu jednak nie udawało się znaleźć dobrego kwantowomechanicznego odpowiednika pola.

Trudno było przejść nad tą luką do porządku dziennego, ponieważ cząstki (opisywane przez mechanikę kwantową) mogą oddziaływać z polami (których mechanika kwantowa nie potrafiła opisać). To tak, jak gdyby ktoś próbował ustalić, jak poruszają się planety Układu Słonecznego, znając zasady dynamiki Newtona (opisujące, jak ciała poruszają się pod wpływem działania sił), ale nie mając pojęcia o prawie powszechnego ciążenia (mówiącym o tym, czym są te siły).

Był jeszcze jeden powód, dla którego fizycy chcieli znaleźć model pól, a nie tylko cząstek. Chodzi o to, że za sprawą dualizmu korpuskularno-falowego oba te pojęcia są ze sobą ściśle związane. Cząstka jest w zasadzie ściśle zwiniętym fragmentem pola. Pole jest natomiast morzem ciasno upakowanych cząstek. Pojęcia te są nierozdzielne. Niestety, używane wówczas metody opierały się na założeniu, że cząstki są maleńkimi punktami, i nie dawały się rozszerzyć w żaden sensowny sposób, tak by mogły opisywać również pola. Nie można po prostu połączyć ze sobą dużej liczby cząstek i nazwać tego polem, ponieważ cząstki *oddziałują* ze sobą.

Wyobraźmy sobie tłum ludzi na… no właśnie, na polu. Powiedzmy, że przyjechali tam na koncert rockowy. Taki tłum oglądany z pokładu helikoptera przypomina ciecz przelewającą się po powierzchni pola – czasem dosłownie, na przykład na festiwalu w Glastonbury, słynnym z tego, że odbywa się w morzu błota. Na powierzchni ziemi wyraźnie natomiast widać, że ta ciecz składa się w istocie z kłębiącej się masy pojedynczych cząstek – ludzi – albo z gęstych skupisk, takich jak grupki znajomych, którzy przechadzają się razem, tworząc nierozdzielne gromadki, czy też z grup obcych sobie osób, podążających w tę samą stronę, na przykład do baru. Nie można jednak przedstawić dokładnego modelu takiego tłumu, dodając po prostu kolejne osoby zachowujące się tak, jak gdyby były zupełnie same. Gdy jedna grupa podąża w kierunku baru, odcina drogę innej gromadce. Obie grupy wpadają na siebie i dochodzi do przepychanek. Skuteczna kwantowa teoria pola musi być opisem takiej właśnie sytuacji, z tym że ludzi należy zastąpić zlokalizowanymi kwantowymi funkcjami falowymi.

Pod koniec lat dwudziestych tego typu rozumowanie przekonało fizyków, że chociaż jest to niezwykle trudne, mechanikę kwantową

trzeba rozszerzyć w taki sposób, by obejmowała nie tylko cząstki, ale i pola. Naturalnym punktem wyjścia dla takich prac było pole elektromagnetyczne. W jakiś sposób trzeba było dokonać kwantyzacji elektrycznych i magnetycznych składników tego pola, czyli zapisać je z wykorzystaniem formalizmu mechaniki kwantowej. Wymagany do tego zapis matematyczny był mało znany i nie przypominał opisu fizycznego. Wielkości wyznaczane w pomiarach nie występowały w tym wypadku w równaniach w postaci starych, dobrych liczb. Teraz odpowiadały im operatory w przestrzeni Hilberta – matematyczne reguły przekształcania fal. Operatory te nie spełniały tradycyjnych założeń mechaniki klasycznej. Jeśli pomnożymy przez siebie dwie liczby, to wynik będzie taki sam, bez względu na to, którą liczbę zapiszemy jako pierwszą. Na przykład 2×3 jest równe 3×2. Własność ta, nazywana przemiennością, nie jest spełniona dla wielu par operatorów, podobnie jak nałożenie skarpetek, a potem butów, nie jest tym samym co nałożenie butów, a potem skarpetek. Liczby są stworzeniami biernymi, natomiast operatory są aktywne. Działanie wykonane jako pierwsze przygotowuje grunt dla następnego.

Przemienność to bardzo wygodna własność matematyczna. Jej brak bywa kłopotliwy i właśnie między innymi z tego powodu kwantowanie pola jest tak trudne. Niemniej czasami udaje się tego dokonać. Kwantyzację pola elektromagnetycznego przeprowadzono w kilku etapach. Pierwszym krokiem była teoria elektronu przedstawiona przez Diraca w 1928 roku, a ostatnim – prace Shin'ichirō Tomonagi, Juliana Schwingera, Richarda Feynmana i Freemana Dysona wykonane na przełomie lat czterdziestych i pięćdziesiątych. W efekcie powstała teoria znana jako elektrodynamika kwantowa.

Wydawało się, że wykorzystane w tych pracach podejście powinno zadziałać również w bardziej ogólnym przypadku. Podstawowy pomysł opierał się na ideach sięgających swymi korzeniami osiągnięć Newtona. W trakcie prac nad rozwiązywaniem równań praw Newtona matematycy odkryli kilka użytecznych, ogólnych sztuczek, które nazywamy zasadami zachowania. Chodzi o to, że w trakcie ruchu układu mas niektóre wielkości nie ulegają zmianie. Najbardziej

znaną z nich jest energia, która występuje w dwóch odmianach: potencjalnej i kinetycznej. Energia kinetyczna ma związek z prędkością danego ciała, natomiast potencjalna odpowiada pracy, jaką mogą wykonać siły. Gdy zepchniemy kamień ze szczytu urwiska, zamieni on swoją energię potencjalną, wynikającą z grawitacji, na energię kinetyczną – mówiąc wprost, będzie spadał z coraz większą prędkością. Inną zachowywaną wielkością jest pęd, będący iloczynem masy i prędkości, oraz moment pędu, który jest związany z szybkością ruchu obrotowego danego ciała. Te zachowywane wielkości łączą ze sobą różne zmienne używane w opisie układu i tym samym zmniejszają ich liczbę. Jest to bardzo pomocne podczas rozwiązywania równań, o czym przekonaliśmy się, analizując problem dwóch ciał w rozdziale 8.

Na początku XX wieku uczonym udało się zrozumieć, z czego wynikają zasady zachowania. Emmy Noether udowodniła, że każda zachowywana wielkość odpowiada ciągłej grupie symetrii równań. Symetria jest przekształceniem matematycznym niezmieniającym równań i wszystkie symetrie tworzą grupę, której działanie można zdefiniować jako: „przeprowadź najpierw jedno, a potem drugie przekształcenie". Grupa ciągła jest grupą symetrii określonych jedną liczbą rzeczywistą. Przekształceniem jest na przykład obrót wokół danej osi, kąt obrotu zaś może być liczbą rzeczywistą, a zatem obroty – o dowolny kąt – wokół danej osi tworzą rodzinę przekształceń ciągłych. W tym przykładzie zachowywaną wielkością jest moment pędu. Podobnie pęd jest wielkością zachowywaną, która jest związana z rodziną przesunięć w danym kierunku. A co z energią? Energia jest wielkością zachowywaną wynikającą z symetrii czasowych – z tego, że równania w każdej chwili są takie same.

Fizycy podejmujący próbę unifikacji podstawowych sił natury przekonali się, że kluczem do tego są symetrie. Pierwszej unifikacji dokonał Maxwell, łącząc elektryczność i magnetyzm w jedno pole elektromagnetyczne. W swojej pracy nie rozważał symetrii, ale szybko stało się jasne, że jego równania mają niezwykły rodzaj

symetrii, którego wcześniej nie zauważono: symetrię cechowania. Wydawało się, że może ona być strategicznym punktem zaczepienia, dzięki któremu uda się opracować bardziej ogólne kwantowe teorie pola.

Obroty i przesunięcia są symetriami globalnymi – przekształcają tak samo całą przestrzeń i czas. Obrót wokół wybranej osi przesuwa każdy punkt przestrzeni o taki sam kąt. Symetrie cechowania są inne – są to symetrie lokalne, które mogą się zmieniać w różnych punktach przestrzeni. Dla elektromagnetyzmu takimi symetriami lokalnymi są zmiany fazy. Pole elektromagnetyczne drgające w określonym miejscu ma swoją amplitudę (wielkość drgań) i fazę (moment, w którym osiąga wartość maksymalną). Jeśli w rozwiązaniu równań pola Maxwella zmienimy fazę w każdym punkcie, to otrzymamy inne rozwiązanie, pod warunkiem że zrównoważymy te zmiany, wprowadzając modyfikacje do opisu pola włączające do niego miejscowy ładunek elektromagnetyczny.

Symetrie cechowania wprowadził do fizyki Hermann Weyl przy okazji nieudanej próby dokonania unifikacji elektromagnetyzmu z ogólną teorią względności. Innymi słowy, chciał połączyć siłę elektromagnetyczną z grawitacyjną. Dziwna nazwa tych symetrii jest wynikiem pewnego nieporozumienia: Weyl sądził, że poprawne symetrie lokalne powinny być zmianami skali przestrzennej, czyli „cechowaniem". Ten pomysł się nie sprawdził, ale w wyniku analizy zapisu matematycznego mechaniki kwantowej Władimir Fok i Fritz London wprowadzili inny rodzaj symetrii lokalnej. Mechanikę kwantową opisuje się za pomocą liczb zespolonych, a nie tylko rzeczywistych, i faza każdej kwantowej funkcji falowej jest wartością zespoloną. Wspomniane symetrie lokalne obracają fazę o dowolny kąt na płaszczyźnie zespolonej. Z abstrakcyjnego punktu widzenia taka grupa symetrii składa się ze wszystkich obrotów, ale w wypadku współrzędnych zespolonych są to „przekształcenia unitarne" (U) w przestrzeni o jednym wymiarze zespolonym (1), dlatego grupę utworzoną przez te symetrie oznacza się jako U(1). Takie ujęcie nie jest tylko abstrakcyjną grą matematyczną – dzięki niemu fizykom udało się zapisać, a następnie rozwiązać równania opisujące naładowane cząstki kwantowe przemieszczające się w polu

elektromagnetycznym. Przyjąwszy taki właśnie punkt widzenia, Tomonaga, Schwinger, Feynman i Dyson zdołali opracować pierwszą relatywistyczną kwantową teorię pola elektromagnetyzmu: elektrodynamikę kwantową. Symetria związana z grupą cechowania U(1) odegrała podstawową rolę w ich pracy.

Następnym krokiem było połączenie elektrodynamiki kwantowej ze słabym oddziaływaniem jądrowym. Dokonali tego Abdus Salam, Sheldon Glashow, Steven Weinberg i inni w latach sześćdziesiątych. Oprócz pola elektromagnetycznego z jego symetrią cechowania U(1) uczeni ci wprowadzili pola związane z czterema cząstkami elementarnymi, tak zwanymi bozonami W^+, W^0, W^- oraz B^0 (bozony W nazywa się również wuonami). Symetrie cechowania tych pól, które w zasadzie są obrotami kombinacji cząstek prowadzącymi do powstania innych kombinacji, tworzą następną grupę, którą oznaczamy symbolem SU(2) – są to przekształcenia unitarne (U) w dwuwymiarowej przestrzeni zespolonej (2), które określa się dodatkowo jako specjalne (S), co sprowadza się do pewnego prostego warunku. Połączoną grupą cechowania jest zatem U(1) × SU(2), gdzie symbol × oznacza, że obie grupy działają niezależnie na oba pola. Tak powstała teoria elektrosłaba, które wymagała wprowadzenia skomplikowanych nowych rozwiązań matematycznych. Grupa U(1) elektrodynamiki kwantowej jest przemienna – przeprowadzenie po kolei dwóch przekształceń symetrycznych daje zawsze taki sam wynik, bez względu na to, które przekształcenie wykonamy jako pierwsze. Ta wygodna właściwość upraszcza opis matematyczny, ale niestety nie jest spełniona w wypadku grupy SU(2). Teoria elektrosłaba była pierwszym przykładem zastosowania nieprzemiennej teorii cechowania.

Gdy rozważamy wewnętrzną strukturę cząstek takich jak protony i neutrony, musimy uwzględnić rolę, jaką odgrywa w nich silne oddziaływanie jądrowe. Do wielkiego przełomu w tym zakresie badań doszło po zauważeniu dziwnej matematycznej prawidłowości związanej z pewną klasą cząstek, które nazywamy hadronami. Prawidłowość ta, znana pod nazwą „ośmiokrotnej ścieżki”, doprowadziła do powstania teorii chromodynamiki kwantowej, która

przewiduje istnienie ukrytych cząstek zwanych kwarkami. W chromodynamice kwantowej kwarki są podstawowymi elementami, z których zbudowane są liczne i różnorodne hadrony.

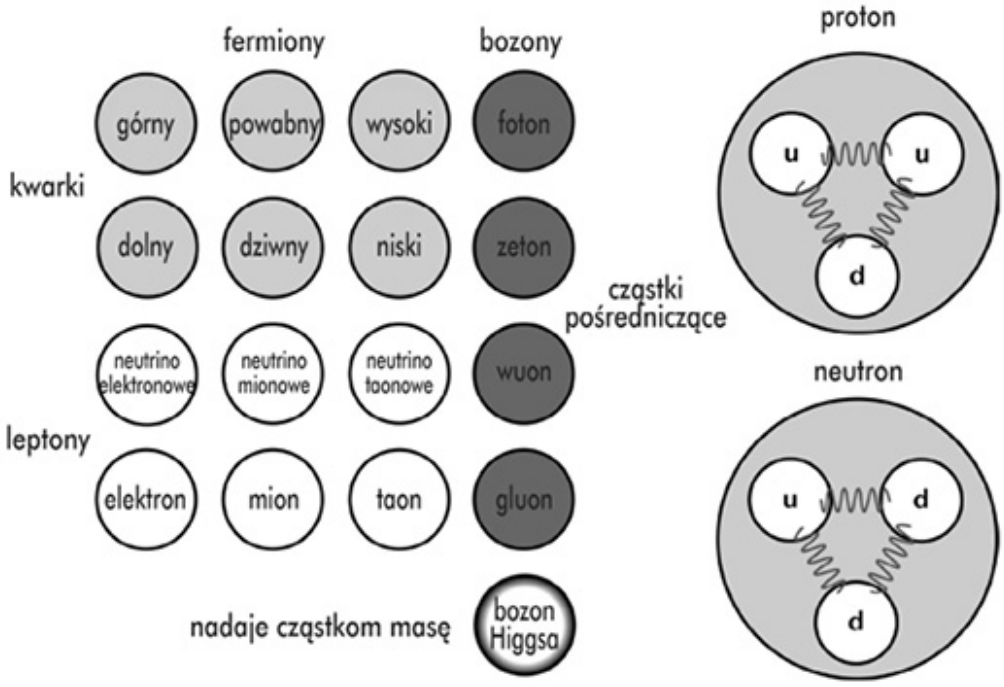
Model Standardowy zakłada, że wszystko, co istnieje we Wszechświecie, zbudowane jest z szesnastu naprawdę podstawowych cząstek, których istnienie udało się potwierdzić w doświadczeniach przeprowadzanych w akceleratorach. Do tego dochodzi jeszcze siedemnasta cząstka, której poszukiwania trwają obecnie w Wielkim Zderzaczu Hadronów. Spośród wszystkich cząstek, które znał Rutherford, tylko dwie okazały się naprawdę podstawowe: są to elektron i foton. Protony i neutrony są zbudowane z kwarków. Taką dziwną nazwę dla nowych cząstek wprowadził Murray Gell-Mann – początkowo chciał, żeby rymowała się z angielskim słowem *cork* (korek). Gdy zastanawiał się nad nią, wpadła mu w ręce książka *Finneganów tren* Jamesa Joyce'a, w której znalazł taki fragment:

Niech kwarki trzy ma Mark!
Choć nieszczególnie nośny ma bark,
Choć jest niezbyt celny i wart niewiele mark.[83]

Słowo kwark od razu przypadło mu do gustu. Wprawdzie nie rymowało się z *cork*, ale miało szlachetny, literacki rodowód.

Model Standardowy przewiduje istnienie sześciu kwarków pogrupowanych w pary. Mają one niezwykłe nazwy: górny i dolny, powabny i dziwny oraz wysoki i niski. Istnieje również sześć leptonów, także pogrupowanych w pary – są to: elektron, mion i taon oraz związane z nimi neutrina. Te dwanaście cząstek to fermiony, nazwane tak od nazwiska Enrica Fermiego. Cząstki utrzymują się razem za sprawą czterech sił: grawitacji, elektromagnetyzmu, silnego oddziaływania jądrowego oraz słabego oddziaływania jądrowego. Musimy tu pominąć grawitację, ponieważ nie udało się jej jeszcze w pełni uzgodnić z kwantowym opisem świata, pozostają nam więc trzy siły. W fizyce cząstek siły powstają w wyniku wymiany cząstek, które „przenoszą" oddziaływania lub, inaczej mówiąc, „pośredniczą" w ich przekazywaniu. Zwykle stosuje

się porównanie do dwóch graczy tenisowych, którzy nie oddalają się zbytnio od siebie, ponieważ są skupieni na krążącej między nimi piłce. Foton jest cząstką pośredniczącą oddziaływania elektromagnetycznego, zetony (bozony Z^0) i wuony pośredniczą w przekazywaniu słabego oddziaływania jądrowego, a gluony przenoszą silne oddziaływanie jądrowe. Ściśle rzecz biorąc, gluony przenoszą siłę koloru, która utrzymuje razem kwarki, a silne oddziaływanie jądrowe jest tym, co obserwujemy w wyniku działania tej siły. Proton składa się z dwóch kwarków górnych i jednego dolnego, natomiast neutron tworzą dwa kwarki górne i jeden dolny. W każdej z tych cząstek kwarki związane są ze sobą dzięki gluonom. Te cztery rodzaje cząstek pośredniczących tworzą grupę bozonów, nazwaną tak od nazwiska Satyendry Bosego. Rozróżnienie między fermionami i bozonami jest bardzo ważne – cząstki te mają różne własności. Na rycinie 44 po lewej stronie pokazano tak zdefiniowany zbiór cząstek podstawowych. Po prawej stronie tej ryciny możemy zobaczyć, w jaki sposób kwarki tworzą protony i neutrony.



Ryc. 44. Po lewej: Siedemnaście cząstek Modelu Standardowego. Po prawej: Protony i neutrony zbudowane są z kwarków. Po prawej na górze: Proton = dwa kwarki górne (u) + jeden kwark dolny (d). Po prawej na dole: Neutron = jeden kwark górny (u) + dwa kwarki dolne (d).

Dopełnieniem tego obrazu jest bozon Higgsa, który jest cząstką wyjaśniającą, dlaczego pozostałe cząstki Modelu Standardowego mają niezerową masę. Nazwano go tak od nazwiska Petera Higgsa, jednego z fizyków, którzy wysunęli ideę działania takiego mechanizmu. Pozostałymi byli: Philip Anderson, François Englert, Robert Brout, Gerald Guralnik, Carl Hagen i Thomas Kibble. Bozon Higgsa jest cząstką będącą ucieleśnieniem hipotetycznego pola kwantowego – pola Higgsa – o pewnej niezwykłej, ale kluczowej własności: otóż w próżni pole to ma niezerową wartość. Pole Higgsa wpływa na pozostałe 16 cząstek i w efekcie zachowują się one tak, jak gdyby miały masę.

W 1993 roku w odpowiedzi na konkurs ogłoszony przez brytyjskiego ministra nauki Williama Waldegrave'a David Miller zaproponował barwne porównanie działania mechanizmu Higgsa do przyjęcia koktajlowego. Wyobraźmy sobie salę wypełnioną równomiernie przez uczestników przyjęcia w momencie, gdy wkracza do niej gość honorowy – była pani premier. Natychmiast wszyscy zaczynają się wokół niej tłoczyć. Pani premier idzie na drugi koniec sali i po drodze wokół niej kłębi się ciągle zmieniająca się grupka gości. W efekcie poruszająca się gromadka ma dodatkową masę i pani premier nie można tak łatwo zatrzymać. Tak właśnie działa mechanizm Higgsa. Wyobraźmy sobie teraz, że w pokoju rozchodzi się jakaś najnowsza plotka i ludzie zbijają się w grupki, żeby usłyszeć nowinę. Taka grupa jest bozonem Higgsa. Na zakończenie Miller dodał jeszcze: „Istnieniu mechanizmu Higgsa i pola Higgsa wypełniającego cały Wszechświat wcale nie musi towarzyszyć bozon Higgsa. O tym, czy tak jest, dowiemy się z doświadczeń, które uczeni przeprowadzą w zderzaczach atomów następnej generacji". Obecnie

wydaje się, że badaczom udało się wyjaśnić kwestię bozonu Higgsa, jednak koncepcja pola Higgsa wymaga jeszcze dalszych prac.

Chromodynamika kwantowa jest kolejną teorią cechowania, tym razem z grupą cechowania SU(3). Jak wynika z tego zapisu symbolicznego, przekształcenia działają teraz w trójwymiarowej przestrzeni zespolonej. Teoria ta dokonuje unifikacji elektromagnetyzmu oraz oddziaływania słabego i silnego. Zakłada istnienie trzech pól kwantowych, po jednym dla każdej z sił, o symetriach cechowania, odpowiednio, U(1), SU(2) i SU(3). Po ich połączeniu uzyskujemy Model Standardowy z grupą cechowania U(1) × SU(2) × SU(3). Ściśle rzecz biorąc, symetrie SU(2) i SU(3) są przybliżone – uważa się, że stają się dokładne w bardzo wysokich energiach. Zatem ich wpływ na cząstki tworzące nasz świat odpowiada „złamanym" symetriom, czyli śladom struktury, które pozostają, gdy układ o symetrii doskonałej zostaje poddany niewielkim zaburzeniom.

Wszystkie trzy grupy zawierają rodziny symetrii ciągłych: jedną w grupie U(1), trzy dla SU(2) i osiem odpowiadających grupie SU(3). Wiążą się z nimi różne zachowywane wielkości. Tak jak wcześniej, symetrie mechaniki newtonowskiej odpowiedzialne są za zachowanie energii, pędu i momentu pędu. Wielkościami zachowywanymi w symetrii cechowania U(1) × SU(2) × SU(3) są różne „liczby kwantowe" opisujące cząstki. Są one odpowiednikami takich wielkości jak spin i ładunek, ale opisują kwarki. Fizycy nazywają je ładunkiem koloru, izospinem i hiperładunkiem. W końcu istnieją jeszcze dodatkowe wielkości zachowywane dla grupy U(1) – są to liczby kwantowe sześciu leptonów, takie jak liczba elektronowa, liczba mionowa i liczba taonowa. W efekcie, tak jak to wynika z twierdzenia Noether, symetrie równań Modelu Standardowego wyjaśniają wszystkie podstawowe zmienne fizyczne cząstek elementarnych.

W tej opowieści najważniejsza jest dla nas ogólna strategia i jej wynik. Jeśli chcemy zunifikować teorie fizyczne, musimy znaleźć ich symetrie i właśnie je poddać unifikacji. Następnie należy opracować

odpowiednią teorię o takiej połączonej grupie symetrii. Zabrzmiało to jak coś bardzo prostego, ale w rzeczywistości stanowi niezwykle skomplikowany proces. W taki jednak sposób powstała kwantowa teoria pola opisująca trzy z czterech sił przyrody – poza zakresem jej obowiązywania pozostaje tylko grawitacja.

Twierdzenie Noether nie tylko wyjaśnia główne zmienne fizyczne związane z cząstkami elementarnymi – dzięki niemu udało się nawet znaleźć wiele z podstawowych symetrii. Fizycy zastosowali także proces odwrotny i na podstawie zaobserwowanych i wyznaczonych teoretycznie liczb kwantowych ustalili, jakie symetrie powinien mieć taki model. Zapisali odpowiednie równania mające takie symetrie i przekonali się, że bardzo dobrze oddają one rzeczywistość. W chwili obecnej ten ostatni krok wymaga odpowiedniego dobrania wartości 19 parametrów – liczb, które trzeba wstawić do równań, by można z nich uzyskać przewidywania ilościowe. Dziewięć z nich to masy konkretnych cząstek: sześciu kwarków, elektronu, mionu i taonu. Pozostałe mają bardziej skomplikowane znaczenie – są to na przykład kąty mieszania i sprzężenia fazowe. Siedemnaście z tych parametrów wyznaczono doświadczalnie, lecz wartości dwóch pozostałych nie udało się w ten sposób ustalić. Są to parametry opisujące wciąż hipotetyczne *pole* Higgsa. Obecnie jednak mamy duże szanse na to, że uda się je w końcu zmierzyć, ponieważ fizycy wiedzą już, gdzie należy ich szukać.

Równania wykorzystywane w tych teoriach należą do ogólnej klasy równań teorii pola z cechowaniem, nazywanych równaniami pola Yanga–Millsa. W 1954 roku Chen-Ning Yang i Robert Mills próbowali opracować teorie cechowania wyjaśniające oddziaływanie silne i związane z nim cząstki. Na początku nie mogli sobie poradzić z kwantyzacją pola, ponieważ krok ten wymagał istnienia cząstek o zerowej masie. W 1960 roku Jeffrey Goldstone, Yōichirō Nambu i Giovanni Jona-Lasinio wymyślili sposób na obejście tego problemu: należy wyjść od teorii przewidującej istnienie bezmasowych cząstek, ale potem zmodyfikować ją odpowiednio, łamiąc niektóre z symetrii. Innymi słowy, należało nieznacznie zmienić równania, wprowadzając do nich nowe asymetryczne wyrazy. Po zastosowaniu takiego podejścia w teorii Yanga–Millsa uzyskano równania opisujące bardzo

dobrze zarówno teorię elektrosłabą, jak i chromodynamikę kwantową.

Yang i Mills założyli, że grupa cechowania jest specjalną grupą unitarną. W wypadku cząstek jest to grupa SU(2) albo SU(3), czyli specjalna grupa unitarna o dwóch lub trzech wymiarach zespolonych, ale takie podejście działa również dla dowolnej innej liczby wymiarów. Ich teoria rozprawia się z pewną trudną, ale nieuniknioną matematyczną przeszkodą. Pole elektromagnetyczne jest pod jednym względem zwodniczo proste: jego symetrie cechowania są przemienne. W przeciwieństwie do większości operatorów kwantowych kolejność dokonywania zmian fazy nie wpływa na równania. Fizyków interesowała jednak kwantowa teoria pola cząstek subatomowych. W takiej sytuacji grupa cechowania jest nieprzemienna, co sprawia, że skwantowanie równań staje się znacznie trudniejsze.

Yang i Mills zdołali dokonać odkrycia dzięki temu, że wykorzystali metodę przedstawiania oddziaływań cząstek na diagramach, wprowadzoną przez Richarda Feynmana. Każdy stan kwantowy można uważać za superpozycję niezliczonych oddziaływań cząstek. Nawet próżnia jest kłębowiskiem par cząstek i antycząstek, które pojawiają się na chwilę i zaraz potem znikają. Zwykłe zderzenie dwóch cząstek przekształca się w zdumiewający taniec pojawiających się i zaraz znikających cząstek pośredniczących, które pędzą wte i wewte, nieustannie dzieląc się i łącząc. Cały ten chaos udaje nam się ogarnąć dzięki dwóm cechom. Po pierwsze, równania pola można skwantować dla każdego określonego diagramu Feynmana i wszystkie wkłady wnoszone przez te diagramy można do siebie dodać, otrzymując efekt wypadkowy będący pełnym oddziaływaniem. Po drugie, najbardziej skomplikowane diagramy odpowiadają przypadkom zdarzającym się niezwykle rzadko i dlatego nie wnoszą one zbyt wiele do ogólnego wyniku. Mimo to istnieje pewien poważny problem. Suma taka, jeśli zinterpretujemy ją dosłownie, jest nieskończona. Yang i Mills znaleźli sposób na dokonanie „renormalizacji” obliczeń, dzięki czemu zdołali pozbyć się nieskończoności różnych wyrazów, które tak naprawdę nie mają znaczenia. Pozostała tylko skończona suma i jej wartość bardzo

dobrze zgadza się z rzeczywistością. Technika ta wydawała się początkowo zupełnie tajemnicza, ale obecnie rozumiemy już, dlaczego działa.

W latach siedemdziesiątych do tych prac włączyli się matematycy i Michael Atiyah uogólnił teorię Yanga–Millsa na szeroką klasę grup cechowania. Matematycy i fizycy zaczęli się wzajemnie inspirować. Prace Edwarda Wittena i Nathana Seiberga dotyczące topologicznych kwantowych teorii pola zaowocowały wprowadzeniem pojęcia supersymetrii. Zgodnie z tą koncepcją wszystkie znane cząstki mają swoje nowe, „supersymetryczne" odpowiedniki: elektronom odpowiadają selektrony, a kwarkom – skwarki. Pozwoliło to uprościć opis matematyczny i wysunąć konkretne fizyczne przewidywania. Tych nowych cząstek nie udało się niestety jeszcze wykryć, chociaż niektóre z nich powinny już chyba dawno pojawić się w doświadczeniach przeprowadzanych w Wielkim Zderzaczu Hadronów. Wartość matematyczna tych idei jest niepodważalna, ale ich bezpośrednie fizyczne znaczenie wciąż pozostaje niepewne. Koncepcje te stawiają jednak teorię Yanga–Millsa w interesującym świetle.

Kwantowa teoria pola jest jednym z najprężniej rozwijających się działów fizyki matematycznej, dlatego Instytut Claya chciał umieścić jedno zagadnienie z tej dziedziny na liście problemów milenijnych. Hipoteza luki masowej jest ważnym elementem tego niezwykle bogatego obszaru badań. Dotyczy ona głębokiego matematycznego problemu związanego z fizyką cząstek. Możliwość zastosowania pól Yanga–Millsa do opisu cząstek elementarnych za pomocą silnego oddziaływania jądrowego zależy od istnienia określonej własności kwantowomechanicznej zwanej luką masową. W teorii względności cząstka poruszająca się z prędkością światła osiąga nieskończoną masę, z wyjątkiem sytuacji, gdy jej masa jest zerowa. Dzięki istnieniu luki masowej cząstki kwantowe mogą mieć skończone niezerowe masy, mimo że związane z nimi fale klasyczne przemieszczają się z prędkością światła. Gdy istnieje luka masowa, dowolny stan niebędący próżnią ma energię wyższą od energii próżni przynajmniej o pewną stałą wartość. Innymi słowy, istnieje pewna niezerowa wartość minimalna masy cząstki.

Doświadczenia potwierdzają istnienie luki masowej, a symulacje komputerowe oparte na odpowiednich równaniach sugerują, że hipoteza luki masowej powinna być poprawna. Nie możemy jednak po prostu przyjąć, że model taki odpowiada rzeczywistości, a następnie posłużyć się rzeczywistością do sprawdzenia poprawności cech matematycznych modelu, ponieważ wpadniemy wtedy w błędne koło. Musimy zrozumieć ten problem na poziomie teoretycznym. Jednym z kluczowych kroków byłoby przedstawienie ścisłego dowodu istnienia kwantowych wersji teorii Yanga–Millsa. Jej klasyczną (niekwantową) wersję rozumiemy już obecnie zupełnie dobrze, ale w wypadku jej kwantowego odpowiednika musimy się zmagać z problemem renormalizacji, czyli owymi kłopotliwymi nieskończonościami, których trzeba się jakoś pozbyć za pomocą matematycznych sztuczek.

Punktem wyjścia jednego z obiecujących podejść jest zamiana ciągłej przestrzeni na dyskretną sieć i zapisanie sieciowego odpowiednika równania Yanga–Millsa. Główny problem polega wtedy na pokazaniu, że w miarę jak sieć staje się coraz gęstsza, upodabniając się do kontinuum, dyskretna wersja równania Yanga–Millsa zbiega do dobrze określonego matematycznego obiektu. Niektóre potrzebne do tego własności matematyczne można określić na podstawie odpowiednich cech fizycznych i być może uda się udowodnić, że odpowiednia teoria kwantowa Yanga–Millsa istnieje, jeśli cechy te da się opisać w sposób ścisły. Wyjaśnienie hipotezy luki masowej wymaga bardziej szczegółowego zrozumienia, w jaki sposób takie teorie sieciowe przybliżają hipotetyczną teorię Yanga–Millsa. Zatem istnienie takiej teorii jest ściśle związane z prawdziwością hipotezy luki masowej.

Na tym jednak prace utknęły. W 2004 roku Michael Douglas przygotował raport o stanie prac nad tym problemem, w którym stwierdził: „O ile mi wiadomo, od kilku lat nie udało się dokonać żadnego przełomu na tym froncie badań. Chociaż dokonano postępu w pracach nad teoriami pola o mniejszej liczbie wymiarów, nie słyszałem o żadnych znaczących osiągnięciach w badaniach nad uzyskaniem matematycznie ścisłej konstrukcji kwantowej teorii

Yanga–Millsa”. Wydaje się, że stwierdzenie to nie straciło nic na swojej aktualności.

Większych postępów udało się natomiast dokonać w badaniach nad pewnymi zagadnieniami związanymi z tym problemem, które mogą rzucić nań nowe światło. Niektóre szczególne kwantowe teorie pola, znane jako dwuwymiarowe modele sigma, są łatwiejsze do analizy matematycznej i dla jednego takiego modelu udało się potwierdzić istnienie luki masowej. Supersymetryczne kwantowe teorie pola, obejmujące hipotetycznych superpartnerów zwyczajnych cząstek elementarnych, mają wygodne cechy matematyczne, które w istocie sprawiają, że renormalizacja przestaje być konieczna. Fizycy, między innymi Edward Witten, pracują nad rozwiązaniem analogicznych równań dla przypadku supersymetrycznego. Mają oni nadzieję, że niektóre metody wykorzystane w tych pracach mogą naprowadzić uczonych na nowe sposoby umożliwiające uporanie się z tym problemem w jego pierwotnej postaci. Jednak bez względu na to, jakie mogą z tych prac wyniknąć wnioski fizyczne i jak zostanie ostatecznie rozstrzygnięta kwestia hipotezy luki masowej, wiele z tych działań już teraz wzbogaciło matematykę o ważne nowe pojęcia i narzędzia.

82 Zob.: Ian Stewart, *17 równań, które zmieniły świat*, przeł. Julia Szajkowska, Prószyński i S-ka, Warszawa 2013, rozdział 14.

83 Przekład Krzysztofa Bartnickiego (przyp. red.).

14. Diofantyczne marzenia

Hipoteza Birchа i Swinnertona-Dyera

W rozdziale 7 mówiliśmy o *Arytmetyce* Diofantosa i wspomnieliśmy wówczas, że sześć spośród trzynastu ksiąg tego dzieła przetrwało w greckich odpisach. Około 400 roku n.e., gdy starożytna cywilizacja grecka zaczęła chylić się ku upadkowi, misję dalszego rozwoju matematyki wzięły na siebie Arabia, Chiny i Indie. Arabscy uczeni przetłumaczyli wiele klasycznych dzieł greckich i nierzadko właśnie takie przekłady są naszymi głównymi źródłami historycznymi. Świat arabski znał *Arytmetykę* i tamtejsi uczeni rozwijali dalej zapoczątkowane w niej dzieło. W 1968 roku odkryto cztery arabskie rękopisy, które mogą być przekładem zaginionych ksiąg *Arytmetyki*.

Pod koniec X wieku perski matematyk Abu Bakr al-Karadżi zadał pytanie, które równie dobrze mógłby postawić Diofantos. Jakie liczby naturalne mogą być wspólną różnicą par trzech kwadratów liczb wymiernych tworzących ciąg arytmetyczny? Na przykład kwadraty liczb naturalnych 1, 25 i 49 mają wspólną różnicę wynoszącą 24. Innymi słowy, 1 + 24 = 25 oraz 25 + 24 = 49. Al-Karadżi żył w latach 953–1029, mógł więc mieć dostęp do egzemplarza *Arytmetyki*, choć najwcześniejszego znanego przekładu tego dzieła dokonał Abu al-Wafa w 998 roku. Leonard Dickson, autor trzytomowej historii teorii liczb, sugeruje, że problem ten mógł się pojawić jeszcze przed 972 rokiem w jakimś anonimowym arabskim rękopisie.

Używając języka algebry, możemy sformułować to pytanie następująco: dla jakich liczb naturalnych *d* istnieje liczba wymierna *x* taka, że *x* – *d*, *x* oraz *x* + *d* są kwadratami? Problem ten można przekształcić do innej równoważnej postaci, choć wcale nie jest to takie oczywiste: jakie liczby naturalne mogą być polem powierzchni trójkąta prostokątnego o długościach boków wyrażających się liczbami wymiernymi? To znaczy jeśli *a*, *b* i *c* są liczbami wymiernymi oraz $a^2 + b^2 = c^2$, to jakie rozwiązania całkowite wyrażają się wzorem $ab/2$? Liczby naturalne spełniające te równoważne warunki nazywamy liczbami przystającymi. Określenie to nie ma żadnego związku z innymi zastosowaniami przymiotnika „przystający" w matematyce, co sprawia, że dla współczesnego czytelnika jest ono nieco mylące. W dalszej części rozdziału wyjaśnimy, skąd się wzięła taka nazwa.

Niektóre liczby nie są przystające. Można na przykład udowodnić, że 1, 2, 3 i 4 nie są przystające. Inne – takie jak 5, 6 i 7 – są przystające. Trójkąt 3–4–5 ma pole 3 × 4/2 = 6, co dowodzi, że 6 jest liczbą przystającą. Aby udowodnić, że 7 jest liczbą przystającą, wystarczy zauważyć, iż liczby $(24/5)^2$, $(25/12)^2$ i $(337/60)^2$ mają wspólną różnicę wynoszącą 7. Do liczby 5 powrócimy za chwilę. Postępując tak z kolejnymi liczbami, możemy przedstawić długą listę liczb przystających, ale nie dowiemy się w ten sposób nic ciekawego o ich naturze. Nawet najdłuższy spis takich przykładów nie pozwoli udowodnić, że konkretna liczba naturalna *nie* jest przystająca. Przez całe stulecia nie było na przykład wiadomo, czy 1 jest liczbą przystającą.

Obecnie wiemy już, że problem ten wykracza daleko poza cokolwiek, co mógłby rozwiązać Diofantos. To pozornie proste pytanie wciąż nie doczekało się pełnej odpowiedzi. Najbliższy jej udzielenia był Jerrold Tunnell, gdy w 1983 roku przedstawił charakterystykę liczb przystających. Z koncepcji Tunnella wynika algorytm pozwalający stwierdzić, czy daną liczbę całkowitą można uzyskać, zliczając liczbę jej przedstawień w postaci dwóch różnych kombinacji kwadratów. Przy odrobinie sprytu takie obliczenia można przeprowadzić dla dość dużych liczb całkowitych. Opis ten ma tylko

jedną wadę – nigdy nie udowodniono, że jest poprawny. Jego poprawność zależy od rozwiązania jednego z problemów milenijnych – hipotezy Bircha i Swinnertona-Dyera. Hipoteza ta opisuje kryterium pozwalające stwierdzić, kiedy krzywa eliptyczna ma skończoną liczbę punktów wymiernych. O tego typu równaniach diofantycznych mówiliśmy już w rozdziale 6 poświęconym hipotezie Mordella i w 7, który traktował o wielkim twierdzeniu Fermata. W tym rozdziale znajdziemy kolejny dowód na to, że równania te odgrywają niezwykle istotną rolę w rozwoju teorii liczb.

W Europie pierwszym uczonym, który zajął się tego typu pytaniami, był Leonardo z Pizy. Leonardo najczęściej kojarzony jest z dziwnym ciągiem liczbowym, który, jak się wydaje, sam wymyślił. Ciąg ten wynika z analizy zadania arytmetycznego dotyczącego potomstwa bardzo nierealistycznych królików. Chodzi o ciąg Fibonacciego

0 1 1 2 3 5 8 13 21 34 55 89…,

w którym każdy wyraz, począwszy od trzeciego, jest sumą dwóch poprzednich. Ojcem Leonarda był urzędnik celny Guglielmo Bonaccio i słynny przydomek oznacza po prostu „syn Bonacciego”. Nie ma żadnych dowodów na to, że używano go za życia Leonarda, i uważa się, że wymyślił go w XIX wieku francuski matematyk Guillaume Libri[84]. Bez względu na to, jak to w istocie było, pozostaje faktem, że liczby Fibonacciego mają wiele fascynujących własności i są powszechnie znane. Pojawiają się nawet w thrillerze Dana Browna *Kod Leonarda da Vinci*.

Leonardo posłużył się liczbami Fibonacciego w podręczniku arytmetyki z 1202 roku, zatytułowanym *Liber Abbaci* (*Księga rachunków*), który miał na celu zwrócić uwagę Europejczyków na nową notację arytmetyczną stosowaną przez Arabów, opierającą się na wykorzystaniu dziesięciu cyfr 0–9, i dowieść jej użyteczności. Idee te dotarły do Europy już wcześniej dzięki książce al-Chuwarizmiego, którą przełożono na łacinę pod tytułem *Algoritmi de numero Indorum* (*Algoritmi o liczbach indyjskich*). Jednak to książka

Leonarda była pierwszą, którą napisano z myślą o wprowadzeniu notacji dziesiętnej w Europie. Znaczna część dzieła została poświęcona praktycznym zadaniom rachunkowym, szczególnie związanym z przeliczaniem walut. Leonardo napisał również inną książkę, która nie jest tak szeroko znana. Pod wieloma względami jest ona europejską kontynuacją *Arytmetyki* Diofantosa – dzieło to nosi tytuł *Liber quadratorum* (*Księga kwadratów*).

Podobnie jak Diofantos, Leonardo przedstawił w niej ogólne metody na konkretnych przykładach. Jeden z nich wykorzystuje pytanie postawione przez al-Karadżiego. W 1225 roku cesarz Fryderyk II odwiedził Pizę. Wiedział, że Leonardo jest słynnym matematykiem, i doszedł do wniosku, że zabawnie byłoby namówić go do uczestnictwa w pojedynku matematycznym. Takie publiczne formy rywalizacji były wówczas dość popularne. Zawodnicy zadawali sobie nawzajem pytania. W skład zespołu cesarza wchodzili Jan z Palermo i mistrz Theodore, a w zespole Leonarda był… Leonardo. Zespół cesarza polecił Leonardowi znaleźć kwadrat, który pozostaje kwadratem, gdy się go powiększy lub pomniejszy o 5. Jak zwykle, rozważano tylko liczby wymierne. Innymi słowy, zespół cesarski chciał przedstawienia dowodu, że 5 jest liczbą przystającą, przez znalezienie konkretnej liczby wymiernej x takiej, że $x - 5$, x oraz $x + 5$ są kwadratami. Nie jest to trywialne – najmniejszym rozwiązaniem jest wartość

$$x = \frac{1681}{144} = \left(\frac{41}{12}\right)^2$$

Wówczas

$$x - 5 = \frac{961}{144} = \left(\frac{31}{12}\right)^2$$

oraz

$$x+5=\frac{2401}{144}=\left(\frac{49}{12}\right)^2$$

Leonardo odkrył rozwiązanie i umieścił je w *Liber quadratorum*. Odpowiedź udało mu się znaleźć dzięki wykorzystaniu ogólnego wzoru związanego ze wzorem Euklidesa i Diofantosa na trójki pitagorejskie. Z tego związku otrzymał trzy kwadraty liczb całkowitych o wspólnej różnicy 720, a mianowicie: 31^2, 41^2 oraz 49^2. Następnie podzielił je przez 12^2 = 144, żeby uzyskać trzy kwadraty o wspólnej różnicy[85] 720/144, czyli 5. Wyrażając to za pomocą trójek pitagorejskich, musimy wziąć trójkąt o bokach 9, 40, 41, którego pole wynosi 180, a po podzieleniu tego pola przez 36 otrzymamy trójkąt o bokach 20/3, 3/2, 41/6. Pole takiego trójkąta wynosi 5.

To właśnie w książce Leonarda znajdziemy łacińskie określenie *congruum* – „przystający, kongruentny" – użyte w odniesieniu do zbioru trzech kwadratów tworzących ciąg arytmetyczny. Później Euler używał słowa *congruere* – „łączyć się". Pierwsze dziesięć liczb przystających i odpowiadające im najprostsze trójki pitagorejskie zamieszczono w tabeli 3. Trudno dostrzec w niej jakąkolwiek prostą prawidłowość.

d	**Trójka pitagorejska**
5	3/2, 20/3 41/6
6	3, 4, 5
7	24/5, 35/12, 337/60
13	780/323, 323/30, 106 921/9690
14	8/3, 63/6, 65/6
15	15/2, 4, 17/2
20	3, 40/3, 41/3
21	7/2, 12, 25/2
22	33/35, 140/3, 4901/105
23	80 155/20 748, 41 496/3485, 905 141 617/72 306 780

Tabela 3. Pierwsze dziesięć liczb przystających i odpowiadające im trójki pitagorejskie

Pierwszych odkryć związanych z tym pytaniem dokonali islamscy matematycy, którzy dowiedli, że liczby 5, 6, 14, 15, 21, 30, 34, 65, 70, 110, 154 i 190 są przystające, podobnie jak 18 innych większych liczb. Do tego zbioru Leonardo, Angelo Genocchi (1855) i André Gérardin (1915) dołożyli jeszcze liczby 7, 22, 41, 69, 77 i 43 oraz inne liczby mniejsze od 1000. W 1225 roku Leonardo stwierdził, że 1 nie jest liczbą przystającą, ale nie przedstawił na to żadnego dowodu. Dopiero w 1569 roku Fermat znalazł potwierdzający to dowód. Do 1915 roku udało się wyznaczyć wszystkie liczby przystające mniejsze od 100, ale prace szły bardzo opornie i w 1980 roku w wypadku wielu liczb mniejszych od 1000 wciąż jeszcze nie było wiadomo, czy są przystające, czy nie. O tym, jak trudne są to badania, świadczy odkrycie L. Bastiena, że 101 jest liczbą przystającą. Boki odpowiadającego jej trójkąta prostokątnego mają długości:

$$\frac{711\,024\,064\,578\,955\,010\,000}{118\,171\,431\,852\,779\,451\,900}$$

$$\frac{3\,967\,272\,806\,033\,495\,003\,922}{118\,171\,431\,852\,779\,451\,900}$$

$$\frac{4\,030\,484\,925\,899\,520\,003\,922}{118\,171\,431\,852\,779\,451\,900}$$

Bastien znalazł te liczby w 1914 roku, przeprowadziwszy wszystkie obliczenia ręcznie. Do roku 1986, wykorzystując do tego już komputery, Gerhard Kramarz podał wszystkie liczby przystające mniejsze od 2000.

W pewnym momencie zauważono, że inne, choć związane z tym problemem równanie

$$y^2 = x^3 - d^2x$$

ma rozwiązania *x*, *y* będące liczbami naturalnymi wtedy i tylko wtedy, gdy *d* jest przystające[86]. Jest to oczywiste, jeśli zauważymy, że wyrażenie po prawej stronie znaku równości jest iloczynem *x*, *x* – *d* oraz *x* + *d* i jeśli wszystkie te czynniki są kwadratami, to również ich iloczyn musi być kwadratem. Odwrotne stwierdzenie też jest dość oczywiste. Takie sformułowanie problemu umieszcza go w zasięgu technik badawczych używanych w bogatej i prężnie rozwijającej się dziedzinie teorii liczb. Dla dowolnej stałej wartości *d* równanie to wiąże y^2 z wielomianem sześciennym zmiennej *x*, a zatem definiuje krzywą eliptyczną. Dochodzimy tym samym do wniosku, że problem liczb przystających jest szczególnym przypadkiem pytania, na które teoretycy liczb bardzo chcieliby znać odpowiedź: kiedy krzywa eliptyczna ma przynajmniej jeden punkt wymierny? Pytanie to w żadnym wypadku nie jest oczywiste, nawet w przytoczonym tu przykładzie szczególnej krzywej eliptycznej. Na przykład liczba 157 jest przystająca, ale *najprostszy* trójkąt o takim polu powierzchni ma przeciwprostokątną o długości:

$$\frac{2\,244\,035\,177\,043\,369\,699\,245\,575\,130\,906\,674\,863\,160\,948\,472\,041}{8\,912\,332\,268\,928\,859\,588\,025\,535\,178\,967\,163\,570\,016\,480\,830}$$

Zanim ruszymy dalej, powróćmy jeszcze do sztuczki Leonarda, za pomocą której przeszedł od liczby 720 do 5, i zastosujmy ją do przypadku ogólnego. Jeśli pomnożymy dowolną liczbę przystającą *d* przez kwadrat n^2 liczby naturalnej *n*, to otrzymamy inną liczbę przystającą. Wystarczy wziąć dowolną wymierną trójkę pitagorejską odpowiadającą trójkątowi o polu *d* i pomnożyć te liczby przez *n*. Taka operacja odpowiada pomnożeniu pola trójkąta przez n^2. Tak samo będzie, gdy podzielimy te liczby przez *n* – teraz pole trójkąta zostało podzielone przez n^2. W wyniku takiej operacji uzyskamy liczbę całkowitą tylko wtedy, gdy pole ma czynnik będący kwadratem,

zatem poszukując liczb przystających, wystarczy się ograniczyć do liczb bezkwadratowych, czyli takich, które nie mają czynnika będącego kwadratem. Pierwszych kilka liczb bezkwadratowych to:

1 2 3 5 6 7 10 11 13 14 15 17 19.

Możemy już teraz przedstawić kryterium Tunnella. Nieparzysta liczba bezkwadratowa *d* jest przystająca wtedy i tylko wtedy, gdy liczba (dodatnich i ujemnych) rozwiązań całkowitych *x*, *y*, *z* równania

$$2x^2 + y^2 + 8z^2 = d$$

jest dokładnie dwa razy większa od liczby rozwiązań równania

$$2x^2 + y^2 + 32z^2 = d$$

Parzysta liczba bezkwadratowa *d* jest przystająca wtedy i tylko wtedy, gdy liczba rozwiązań całkowitych *x*, *y*, *z* równania

$$8x^2 + 2y^2 + 16z^2 = d$$

jest dokładnie dwa razy większa od liczby rozwiązań równania

$$8x^2 + 2y^2 + 64z^2 = d$$

Kryterium to jest bardziej użyteczne, niż się może wydawać na pierwszy rzut oka. Ponieważ wszystkie współczynniki równań są dodatnie, wartość liczb *x*, *y* i *z* nie może przekroczyć pewnych wielokrotności pierwiastka kwadratowego z *d*. Wynika stąd, że liczba rozwiązań jest skończona i można je znaleźć, prowadząc systematyczne poszukiwania z zastosowaniem kilku wygodnych uproszczeń. Oto pełne obliczenia dla kilku przykładów o małej wartości *d*:

• Gdy $d = 1$, to jedynymi rozwiązaniami pierwszego równania są $x = 0$, $y = \pm 1$, $z = 0$. To samo dotyczy drugiego równania. Zatem oba równania mają po dwa rozwiązania i kryterium nie jest spełnione.
• Gdy $d = 2$, to jedynymi rozwiązaniami pierwszego równania są $x = \pm 1$, $y = 0$, $z = 0$. To samo dotyczy drugiego równania. Zatem oba równania mają po dwa rozwiązania i kryterium nie jest spełnione.
• Gdy $d = 3$, to jedynymi rozwiązaniami pierwszego równania są $x = \pm 1$, $y = \pm 1$, $z = 0$. To samo dotyczy drugiego równania. Zatem oba równania mają po cztery rozwiązania i kryterium nie jest spełnione.
• Gdy $d = 5$ lub 7, to pierwsze równanie nie ma rozwiązań. To samo dotyczy drugiego równania. Ponieważ dwa razy zero jest równe zero, to kryterium jest spełnione.
• Gdy $d = 6$, musimy wykorzystać kryterium dla liczb parzystych. Również w tym wypadku oba równania nie mają rozwiązań, kryterium jest więc spełnione.

Z tych prostych obliczeń wynika, że 1, 2, 3, 4 ($= 2^2 \times 1$) nie są liczbami przystającymi, ale 5, 6 i 7 już tak. Tę analizę można bez trudu ciągnąć dalej i w 2009 roku zespół matematyków zastosował kryterium Tunnella dla liczb z przedziału od jednego do biliona, dzięki czemu znalazł dokładnie 3 148 379 694 liczby przystające. Badacze sprawdzili swoje wyniki, przeprowadzając obliczenia dwukrotnie na różnych komputerach z wykorzystaniem dwóch algorytmów napisanych przez dwie niezależne grupy. Bill Hart i Gonzalo Tornaria użyli do tego komputera Selmer na uniwersytecie w Warwick. Mark Watkins, David Harvey i Robert Bradshaw przeprowadzili swoje obliczenia na komputerze Sage działającym na Uniwersytecie Stanu Waszyngton w Seattle.

Jednak we wszystkich tego rodzaju obliczeniach istnieje pewna luka. Tunnell udowodnił, że jeśli liczba d jest przystająca, to podane przez niego kryterium musi być spełnione. Zatem jeśli warunki te nie są spełnione, to liczba nie jest przystająca. Wynika z tego na przykład, że liczby 1, 2, 3 i 4 nie są przystające. Tunnellowi nie udało się co prawda udowodnić twierdzenia przeciwnego, a mianowicie

tego, że jeśli jakaś liczba spełnia jego kryterium, to musi być liczbą przystającą. A właśnie taki dowód jest potrzebny, byśmy mogli stwierdzić, że liczby 5, 6 i 7 są przystające. W tych konkretnych przykładach potrafimy znaleźć odpowiednie trójki pitagorejskie, ale takie podejście nie daje się zastosować w ogólnym przypadku. Tunnellowi udało się jednak dowieść, że takie odwrotne stwierdzenie wynika z hipotezy Bircha i Swinnertona-Dyera – tyle tylko że jej również nie udało się jeszcze udowodnić.

Podobnie jak kilka innych problemów milenijnych, hipoteza Bircha i Swinnertona-Dyera jest trudna nawet do sformułowania. (Myśleliście, że można dostać milion dolarów za zrobienie czegoś prostego? Taa, a świstak siedzi i zawija je w te sreberka…). Warto jednak trochę się wysilić, ponieważ przy okazji będziemy mogli docenić głębię i znaczenie teorii liczb. Jak można się domyślić z nazwy hipotezy, wysunęło ją dwóch matematyków: Brian Birch i Peter Swinnerton-Dyer. W pełnym brzmieniu hipoteza ta jest dość skomplikowana, ale w swojej istocie dotyczy pewnej podstawowej kwestii związanej z równaniami diofantycznymi, czyli równaniami algebraicznymi, których rozwiązań poszukujemy tylko wśród liczb całkowitych i wymiernych. Pytanie jest proste: kiedy takie równania mają rozwiązania?

W rozdziale 6 poświęconym hipotezie Mordella i w 7 dotyczącym wielkiego twierdzenia Fermata mieliśmy okazję poznać jedne z najwspanialszych gadżetów w całej matematyce – krzywe eliptyczne. Mordell wysunął dość daleko idące, jak na ówczesne czasy przypuszczenie i zaproponował hipotezę, że liczba wymiernych rozwiązań równania algebraicznego dwóch zmiennych zależy od topologii związanej z nim krzywej zespolonej. Jeśli jej genus wynosi 0 – czyli gdy z topologicznego punktu widzenia jest sferą – to rozwiązania można zapisać w postaci pewnego wzoru. Jeśli jej genus wynosi 1 – czyli gdy z topologicznego punktu widzenia jest torusem, co jest równoważne stwierdzeniu, że jest to krzywa eliptyczna – to wszystkie rozwiązania wymierne można uzyskać z odpowiedniej skończonej listy dzięki zastosowaniu naturalnej struktury grupy. Jeśli

jej genus wynosi 2 lub więcej – czyli gdy z topologicznego punktu widzenia jest torusem z g otworami dla $g \geq 2$ – to liczba rozwiązań jest skończona. Jak już mówiliśmy, Faltings udowodnił to niezwykłe twierdzenie w 1983 roku.

Najbardziej uderzającą cechą rozwiązań wymiernych równań krzywych eliptycznych jest to, że tworzą one grupę, co wynika z konstrukcji przedstawionej na rycinie 28 ze s. 161. Tak powstałą strukturę nazywamy grupą Mordella–Weila danej krzywej. Teoretycy liczb bardzo chcieliby móc ją wyliczyć, wymaga to jednak znalezienia zbioru generatorów – rozwiązań wymiernych, z których można wyprowadzić wszystkie pozostałe, stosując wielokrotnie działanie tej grupy. A jeśli to przekracza nasze możliwości, to chcielibyśmy przynajmniej ustalić, jakie są najbardziej podstawowe cechy tej grupy – na przykład dowiedzieć się, jak jest ona duża. Wielu związanych z tym szczegółów wciąż jeszcze nie rozumiemy. Czasami taka grupa jest nieskończona, co prowadzi do uzyskania nieskończenie wielu rozwiązań wymiernych, a czasami skończona i liczba rozwiązań wymiernych też jest ograniczona. Dobrze byłoby, gdybyśmy mogli stwierdzić za każdym razem, z którym przypadkiem mamy do czynienia. Najbardziej chcielibyśmy jednak poznać strukturę abstrakcyjną tej grupy.

Mordell udowodnił, że wszystkie rozwiązania możemy wygenerować ze skończonej listy, co oznacza, że taką grupę da się utworzyć z pewnej skończonej grupy i specjalnej grupy, którą fachowo nazywa się kratą. Krata składa się ze wszystkich list o określonej, skończonej długości, zawierających liczby całkowite. Jeśli długość ta wynosi na przykład 3, to krata składa się ze wszystkich list (m_1, m_2, m_3), których elementami są liczby całkowite. Listy takie dodaje się do siebie w oczywisty sposób:

$$(m_1, m_2, m_3) + (n_1, n_2, n_3) = (m_1 + n_1, m_2 + n_2, m_3 + n_3).$$

Długość listy nazywamy rangą kraty (a z geometrycznego punktu widzenia jest to wymiar kraty). Jeśli ranga wynosi 0, to krata jest skończona. Jeśli ranga jest różna od zera, to krata jest nieskończona.

Aby więc ustalić, ile istnieje rozwiązań, wcale nie musimy znać pełnej struktury kraty. Wystarczy, że będziemy wiedzieli, jaka jest jej ranga. Właśnie tego dotyczy hipoteza Bircha i Swinnertona-Dyera.

W latach sześćdziesiątych, gdy powstawały pierwsze komputery, na uniwersytecie w Cambridge działał jeden z wczesnych modeli pod nazwą EDSAC. Jest to skrót od ang. *electronic delay storage automatic calculator* (automatyczny kalkulator elektroniczny z pamięcią rtęciową). Nazwa świadczy o tym, jak bardzo konstruktorzy komputera byli dumni z jego układu pamięci, którego działanie opierało się na przesyłaniu fal dźwiękowych na drugi koniec rur wypełnionych rtęcią, gdzie sygnał był odbierany i zawracany na początek rury. Całe urządzenie miało wielkość dużej ciężarówki i do dzisiaj dobrze pamiętam, jak w 1963 roku pokazywano mi ten najnowszy cud techniki. Układy komputera składały się z tysięcy zaworów – rur próżniowych. Wzdłuż ścian stały ich długie rzędy – były to rury zapasowe, które trzeba było wstawić do komputera, gdy któraś z zainstalowanych w nim rur wybuchała. A zdarzało się to dosyć często.

Peter Swinnerton-Dyer interesował się diofantycznym aspektem krzywych eliptycznych, a w szczególności chciał zrozumieć, ile będzie rozwiązań, jeśli zastąpimy krzywą jej odpowiednikiem w skończonym ciele p-elementowym, gdzie p jest liczbą pierwszą. Innymi słowy, chciał sprawdzić, co się stanie po zastosowaniu gaussowskiej sztuczki z działaniami „modulo p”. Wyznaczył te wartości za pomocą komputera dla wielu liczb pierwszych i próbował znaleźć w nich jakieś interesujące prawidłowości.

Po pewnym czasie zaczął coś podejrzewać. Opiekun jego pracy doktorskiej, John William Scott („Ian”) Cassels, odnosił się początkowo z dużą rezerwą do tych pomysłów, ale w miarę jak wzrastała liczba danych, był coraz bliższy przyznania, że coś w tym jednak jest. Powiedzmy więc, co wynikało z doświadczeń komputerowych Swinnertona-Dyera. Teoretycy liczb mają standardowe metody pozwalające na zinterpretowanie dowolnego równania zdefiniowanego dla zwyczajnych liczb całkowitych za pomocą arytmetyki modularnej – przypomnijmy sobie „zegar arytmetyczny” modulo 12 z rozdziału 2. Ponieważ w takiej wersji

arytmetyki wciąż obowiązują zwyczajne zasady algebry, każde rozwiązanie pierwotnego równania staje się jakimś rozwiązaniem równania „zredukowanego” do danej wartości modulo. Jako że w tym wypadku używany zbiór liczb jest ograniczony i dobrze znany – w przykładzie z zegarem arytmetycznym jest ich tylko 12 – metodą prób i błędów możemy znaleźć wszystkie rozwiązania. W szczególności można policzyć, ile istnieje rozwiązań dla danej wartości modulo. Z rozwiązań dla każdego modułu wynikają odpowiednie warunki na pierwotne rozwiązania całkowite i czasami można nawet dzięki nim udowodnić, że takie rozwiązania w ogóle istnieją. Dlatego teoretycy liczb niemal odruchowo redukują równania do różnych wartości modulo, a liczby pierwsze są tu szczególnie dobrym wyborem.

Jeśli więc chcemy dowiedzieć się czegoś o krzywej eliptycznej, powinniśmy rozważyć wszystkie liczby pierwsze aż od określonej wartości. Dla każdej z nich możemy określić, ile punktów leży na tej krzywej modulo dana liczba pierwsza. Birch zauważył, że w doświadczeniach komputerowych Swinnertona-Dyera pojawia się interesująca prawidłowość, jeśli wykona się na nich prostą operację. Najpierw należy podzielić liczbę takich punktów przez rozważaną liczbę pierwszą, a następnie wymnożyć wszystkie ułamki, wyznaczone dla wszystkich liczb pierwszych mniejszych lub równych danej liczbie, i narysować na papierze logarytmicznym uzyskane wyniki dla kolejnych liczb pierwszych. Tak wyznaczone punkty wydają się leżeć blisko prostej o nachyleniu odpowiadającym randze badanej krzywej eliptycznej. Na tej podstawie można zaproponować wzór na liczbę rozwiązań związanych z dowolną wartością modulo będącą liczbą pierwszą[87].

Wzór ten nie wywodzi się jednak z teorii liczb – wymaga użycia analizy zespolonej, ulubionej dziedziny dziewiętnastowiecznych matematyków, która jakimś cudem jest dużo bardziej elegancka od staromodnej analizy matematycznej operującej na liczbach rzeczywistych. W rozdziale 9 poświęconym hipotezie Riemanna przekonaliśmy się już, że analiza wyciąga swoje macki we wszystkich kierunkach i że szczególnie zaskakujące i silne powiązania łączą ją

z teorią liczb. Wzór Swinnertona-Dyera stał się podstawą do wysunięcia bardziej szczegółowej hipotezy na temat pewnego rodzaju funkcji zespolonej, o której wspomnieliśmy już w rozdziale 9, zwanej funkcją *L* Dirichleta. Funkcja ta jest odpowiednikiem dla funkcji eliptycznych słynnej funkcji dzeta Riemanna. Obaj matematycy bez wątpienia poszli na całość, ponieważ w tamtym okresie nie było wcale pewne, czy wszystkie krzywe eliptyczne *mają* funkcje *L* Dirichleta. Było to jedynie przypuszczenie poparte wątpliwymi dowodami. Jednak w miarę jak uczeni poznawali ten obszar, ich śmiałe pomysły nabierały coraz większego uzasadnienia. Nie był to desperacki skok w nieznane, ale raczej przejaw cudownie trafnej, daleko idącej matematycznej intuicji. Zamiast wdrapywać się na barki olbrzymów, Birch i Swinnerton-Dyer stanęli na własnych ramionach – niczym olbrzymy unoszące się w powietrzu.

Jedna z podstawowych metod analizy zespolonej polega na wyrażeniu funkcji za pomocą szeregu potęgowego. Szereg taki przypomina wielomian, ale zawiera nieskończenie wiele wyrazów operujących coraz wyższymi potęgami zmiennej, którą w tej dziedzinie matematyki tradycyjnie oznacza się literą *s*. Aby sprawdzić, jak się zachowuje dana funkcja w pobliżu określonego punktu, powiedzmy 1, należy użyć potęg wyrażenia ($s - 1$). Hipoteza Bircha i Swinnertona-Dyera głosi, że jeśli rozwinięcie w szereg potęgowy funkcji *L* Dirichleta w pobliżu 1 ma postać:

$$L(C, s) = c(s - 1)^r + \text{wyrazy rzędu wyższego},$$

gdzie *c* jest niezerową stałą, to ranga krzywej wynosi *r* i odwrotnie. W języku analizy zespolonej stwierdzenie to przyjmuje postać: „*L*(*C*, *s*) ma *r*-krotne miejsce zerowe w punkcie $s = 1$”.

Najważniejsza nie jest tutaj dokładna postać tego wyrażenia, ale fakt, że dla dowolnej krzywej eliptycznej możemy przeprowadzić obliczenia analityczne wykorzystujące związaną z nią funkcję zespoloną, z których dowiemy się, ile dokładnie niezależnych rozwiązań wymiernych musimy znaleźć, by móc wyznaczyć je wszystkie.

Najprostszym chyba sposobem na udowodnienie, że hipoteza Bircha i Swinnertona-Dyera istotnie wnosi coś nowego, jest spostrzeżenie, że największa znana ranga wynosi 28. Innymi słowy, istnieją krzywe eliptyczne mające zbiór 28 rozwiązań wymiernych, z których można wyprowadzić wszystkie pozostałe. Co więcej, takiej możliwości nie daje żaden mniejszy zbiór rozwiązań wymiernych. Chociaż wiadomo, że krzywe o takiej randze istnieją, jak dotąd nie udało się znaleźć żadnego konkretnego przykładu. Największa ranga, dla której potrafimy przedstawić konkretny przykład, wynosi 18. Krzywą o takiej randze znalazł Noam Elkies w 2006 roku. Jej równanie ma postać:

$$y^2 + xy = x^3 - 26\,175\,960\,092\,705\,884\,096\,311\,701\,787\,701\,203\,903\,556\,438\,969\,515x + 51\,069\,381\,476\,131\,486\,489\,742\,177\,100\,373\,772\,089\,779\,103\,253\,890\,567\,848\,326$$

W takim zapisie równanie to nie ma standardowej postaci „y^2 = wielomian sześcienny z x", ale można je przekształcić do takiej formy – wtedy jednak liczby będą jeszcze większe. Przypuszcza się, że ranga może mieć dowolnie dużą wartość, ale nikt jeszcze tego nie udowodnił. O ile nam obecnie wiadomo, nigdy nie przekracza ona ustalonej wartości granicznej.

Większość tego, co potrafimy udowodnić, dotyczy krzywych o randze 0 lub 1. Dla rangi o wartości 0 istnieje skończona liczba rozwiązań wymiernych. Gdy ranga wynosi 1, jedno określone rozwiązanie pozwala uzyskać wszystkie pozostałe, być może z wyłączeniem skończonej liczby wyjątków. Te dwa przypadki obejmują wszystkie krzywe eliptyczne postaci $y^2 = x^3 + px$, gdzie p jest liczbą pierwszą postaci $8k + 5$ (na przykład 13, 29, 37 i tak dalej). Istnieje hipoteza głosząca, że w tych sytuacjach ranga zawsze wynosi 1, a z tego wynika, że istnieje nieskończenie wiele rozwiązań wymiernych. Andrew Bremner i John William Scott Cassels udowodnili, że tak właśnie jest w wypadku wszystkich tego typu liczb pierwszych mniejszych od 1000. Znalezienie rozwiązań, z których można wyprowadzić niemal wszystkie pozostałe, może być dość

trudne, nawet gdy znamy rangę i nie jest ona duża. Bremner i Cassels odkryli na przykład, że gdy $p = 887$, to *najprostsze* rozwiązanie tego rodzaju jest liczbą wymierną postaci:

$$\frac{375\,494\,528\,127\,162\,193\,105\,504\,069\,942\,092\,792\,346\,201}{6\,215\,987\,776\,871\,505\,425\,463\,220\,780\,697\,238\,044\,100}$$

Wiele twierdzeń związanych z hipotezą Bircha i Swinnertona-Dyera udało się udowodnić, zwykle dzięki przyjęciu bardzo skomplikowanych założeń, ale prace na drodze do uzyskania dowodu samej hipotezy postępują dość opornie. W 1976 roku Coates i Wiles odkryli pierwszą wskazówkę przemawiającą za poprawnością tej hipotezy. Udowodnili wówczas, że specjalny rodzaj krzywej eliptycznej ma rangę 0, jeśli funkcja *L* Dirichleta nie zeruje się w 1. Dla takiej krzywej eliptycznej liczba rozwiązań wymiernych równania diofantycznego jest skończona – być może nawet zerowa – i można to wywnioskować z odpowiadającej mu funkcji *L*. Później dokonano jeszcze kilku szczegółowych odkryć, wciąż głównie ograniczających się do rang 0 i 1. W 1990 roku Wiktor Koływagin udowodnił, że hipoteza Bircha i Swinnertona-Dyera jest prawdziwa dla rang 0 i 1.

Wiele bardziej szczegółowych hipotez, za którymi przemawiają liczne wyniki badań komputerowych, wiąże stałą *c* w hipotezie Bircha i Swinnertona-Dyera z różnymi pojęciami z zakresu teorii liczb. Istnieją także odpowiedniki – równie zagadkowe – ciał liczbowych. Wiadomo też, że krzywe eliptyczne w większości – w ściśle określonym znaczeniu tego słowa – mają rangę 0 lub 1. W 2010 roku Manjul Bhargava i Arul Shankar ogłosili, że udało im się udowodnić, iż średnia ranga krzywej eliptycznej jest nie większa niż 7/6. Jeśli ten wynik i kilka innych ogłoszonych niedawno twierdzeń przejdzie pomyślnie weryfikację, to będzie to oznaczało, że hipoteza Bircha i Swinnertona-Dyera jest prawdziwa dla niezerowego odsetka wszystkich krzywych eliptycznych. Są to jednak najprostsze przypadki, które nie obejmują krzywych o najbogatszej strukturze, czyli tych o randze wynoszącej 2 i więcej. Te krzywe stanowią niemal nieprzeniknioną tajemnicę.

84 Zob. angielskie wydanie książki Fibonacciego *Liber quadratorum*: Leonardo Pisano Fibonacci, *The Book of Squares* (*Księga kwadratów*), przełożył i opatrzył komentarzami L.E. Sigler, Academic Press, 1987.

85 Leonardo znalazł całą rodzinę rozwiązań:

$$\left(m^2+n^2\right)^2-4mn\left(m^2-n^2\right)=\left(m^2-2mn-n^2\right)^2$$

$$\left(m^2+n^2\right)^2+4mn\left(m^2-n^2\right)=\left(m^2+2mn-n^2\right)^2$$

gdzie *m* i *n* są nieparzyste. Rolę *d* odgrywa w tym wypadku liczba $4mn\left(m^2-n^2\right)$, natomiast *x* wynosi $\left(m^2+n^2\right)^2$. Jeśli przyjmiemy *m* = 5, *n* = 4, to $4mn\left(m^2-n^2\right)=720$. Co więcej, 720 = 5 × 12^2. Dzieląc *x* przez 12^2, otrzymujemy odpowiedź.

86 Jeśli liczby *x* − *n*, *x* oraz *x* + *n* są kwadratami, to kwadratem musi być również ich iloczyn, który wynosi x^3-n^2x. Zatem równanie $y^2=x^3-n^2x$ ma wymierne rozwiązanie. Co więcej, *y* nie może być równe zero, bo wtedy *x* = *n* i zarówno *x*, jak i 2*x* muszą być kwadratami, a to jest niemożliwe, ponieważ $\sqrt{2}$ jest liczbą niewymierną.

I przeciwnie, jeśli *x* i *y* spełniają to równanie sześcienne, a *y* jest różne od 0, to liczby $a=\left(x^2-n^2\right)/y$, $b=2nx/y$ oraz $c=\left(x^2+n^2\right)/y$ spełniają równania $a^2+b^2=c^2$ oraz $ab/2=n$.

87 Chodzi o wzór

$$\prod_{p\le x}\frac{N_p}{p}\approx C\left(\ln r\right)^r$$

gdzie *r* jest rangą, *C* to stała, natomiast symbol ≈ oznacza, że iloraz wyrażeń po obu jego stronach zmierza do 1, gdy *x* zmierza do nieskończoności.

15. Cykle zespolone

Hipoteza Hodge'a

Niektóre działy matematyki można połączyć dość bezpośrednio z wydarzeniami i problemami życia codziennego. Raczej nie natkniemy się na równanie Naviera–Stokesa w kuchni, ale wszyscy wiemy, czym są ciecze, i instynktownie rozumiemy ich przepływ. Niektóre dziedziny można powiązać z ezoterycznymi pytaniami z pogranicza nauki: być może nie da się zrozumieć kwantowej teorii pola bez doktoratu z fizyki matematycznej, ale wiele można wyjaśnić dzięki porównaniom z elektrycznością i magnetyzmem lub niecodziennymi wyobrażeniami takimi jak „fala prawdopodobieństwa". Czasami pomocne okazują się ilustracje – dobrym tego przykładem jest hipoteza Poincarégo. Pewne koncepcje opierają się jednak wszelkim tego typu próbom przybliżenia trudnych abstrakcyjnych pojęć.

Do tego grona należy hipoteza Hodge'a, sformułowana w 1950 roku przez szkockiego matematyka Williama Hodge'a. Problemem nie jest tu wyjaśnienie dowodu, bo takowego nie ma. Kłopotów nastręcza już samo sformułowanie hipotezy. Oto jej opis zaczerpnięty ze strony internetowej Instytutu Claya z niewielkimi zmianami redakcyjnymi:

> Na dowolnej nieosobliwej zespolonej algebraicznej rozmaitości rzutowej każda klasa Hodge'a jest wymierną kombinacją liniową cykli algebraicznych.

Jak widać, czeka nas mnóstwo pracy. Jedynymi słowami, które wydają się zrozumiałe, są: „na”, „dowolna”, „każda” oraz „jest”. Inne brzmią znajomo: „rozmaitość”, „klasa”, „wymierny”, „cykl”. Jednak obrazy, które przywodzą na myśl – uginająca się od towarów półka w supermarkecie, sala wypełniona uczniami, dające się policzyć korzyści i następujące po sobie pory roku – nie są raczej tym, o co chodziło specjalistom z Instytutu Claya. Pozostałe słowa to już bez wątpienia terminy specjalistyczne. Nie chodzi tu jednak o bezsensowne nadawanie dziwacznych nazw prostym rzeczom. W tym wypadku mamy do czynienia z prostymi nazwami skomplikowanych pojęć. W naszym codziennym języku nie ma żadnych określeń na tego typu idee, niektóre więc sobie pożyczamy, a inne – po prostu wymyślamy.

Spójrzmy jednak na to od pozytywnej strony: nadarza nam się tu niepowtarzalna okazja. Spośród wszystkich zagadnień omawianych w tej książce hipoteza Hodge’a jest chyba najbardziej typowym przykładem prawdziwej matematyki, takiej, jaką uczeni rozwijają w XX i XXI stuleciu. Jeśli podejdziemy do niej w odpowiedni sposób, będziemy mogli się przekonać, jak bardzo zaawansowane pojęciowo są najnowsze badania matematyczne. W porównaniu z matematyką nauczaną w szkole pojęcia te są niczym Mount Everest przy krecim kopczyku.

Czy są to zatem tylko jakieś wydumane, nadęte niedorzeczności zajmujące umysły uczonych, którzy stracili kontakt z rzeczywistością? Skoro nikt poza nimi nie jest w stanie zrozumieć, o co w tym wszystkim chodzi, to po co wydawać pieniądze podatników na zatrudnianie ludzi rozmyślających o takich sprawach? Pozwólcie, że odwrócę to pytanie. Co by było, gdyby każdy *mógł* zrozumieć wszystko, o czym dyskutują matematycy? Czy wtedy chętnie przeznaczalibyśmy na ich badania pieniądze podatników? A może jednak biorą oni pieniądze właśnie za to, że są ekspertami? Gdyby wszystko było tak łatwe i zrozumiałe, że każdy przechodzień spotkany przypadkowo na ulicy z miejsca wiedziałby, o co chodzi, to do czego byliby nam potrzebni matematycy? Gdyby wszyscy potrafili spuścić wodę z instalacji centralnego ogrzewania i naprawić pękniętą rurę, to czy hydraulicy mieliby cokolwiek do roboty?

Nie jestem w stanie wskazać żadnej modnej aplikacji, której działanie opiera się na hipotezie Hodge'a, ale mogę wyjaśnić, dlaczego jest ona ważna w matematyce. Współczesna matematyka jest jednolitą całością, zatem każdy znaczący postęp w dowolnym z jej kluczowych działów przyniesie ostatecznie korzyści dające się przeliczyć na złotówki. Zapewne nie natkniemy się na nią dzisiaj w kuchni, ale jutro… kto wie? Blisko z nią związane pojęcia matematyczne już teraz znalazły zastosowanie w kilku dziedzinach nauki, od fizyki kwantowej i teorii strun po robotykę.

Czasami praktyczne zastosowania nowych odkryć matematycznych są oczywiste niemal od razu. Czasami dopiero po stuleciach. W tym drugim wypadku mogłoby się wydawać, że bardziej opłacalne byłoby poczekać do chwili, gdy dane rozwiązanie będzie potrzebne, i dopiero wtedy zabrać się do roboty pełną parą. Wszystkie problemy matematyczne, które nie mają natychmiastowego, oczywistego zastosowania, powinny być odkładane na bok. Gdybyśmy tak jednak postępowali, zawsze bylibyśmy zapóźnieni, ponieważ matematyka potrzebowała kilkuset lat, by sprostać potrzebom nauk stosowanych. Wcale też nie musi być jasne, która idea jest potrzebna. Czy bylibyśmy zadowoleni, gdyby pracownicy cegielni zabierali się do opracowania metod wyrobu cegieł dopiero wtedy, gdy już zatrudnimy ekipę budowlaną i rozpoczniemy budowę domu? Im bardziej oryginalne jest dane pojęcie matematyczne, tym mniej jest prawdopodobne, że uda się na nie wpaść w toku gorączkowych prac ukierunkowanych na uzyskanie określonego wyniku.

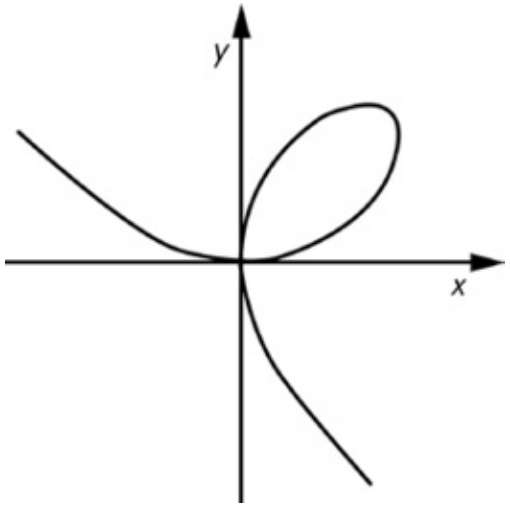
Ze strategicznego punktu widzenia lepszym rozwiązaniem jest pozwolić, żeby niektóre gałęzie matematyki rozwijały się swoim rytmem, bez oczekiwania natychmiastowych korzyści. Nie próbujmy dokonywać arbitralnych wyborów – niech matematyka rozrasta się w sposób organiczny. Matematycy są tani – nie potrzebują kosztownych urządzeń, tak jak fizycy cząstek (Wielki Zderzacz Hadronów – dotychczasowy koszt 7,5 miliarda euro). Odpłacą się nam, nauczając w szkołach i na uczelniach. W świetle tego wszystkiego trudno im odmówić prawa do poświęcenia części swojego czasu na roztrząsanie hipotezy Hogdge'a, jeśli akurat to ich pociąga.

Rozbierzemy teraz sformułowanie hipotezy Hodge'a na części, słowo po słowie. Najprostszym pojęciem jest „rozmaitość algebraiczna". Jest ona naturalną konsekwencją wynikającą z użycia współrzędnych kartezjańskich, które łączą geometrię z algebrą (mówiliśmy o tym w rozdziale 3). Z ich pomocą niewielki zestaw krzywych wprowadzony przez Euklidesa i jego następców – prosta, okrąg, elipsa, parabola, hiperbola – stał się niewyczerpanym rogiem obfitości. Prosta, podstawa całej geometrii euklidesowej, jest zbiorem punktów spełniających odpowiednie równanie, na przykład $y = 3x + 1$. Jeśli zmienimy 3 i 1 na inne liczby, to uzyskamy inne proste. Okręgi wymagają zastosowania równań kwadratowych – podobnie jak elipsy, parabole i hiperbole. W zasadzie wszystko, co daje się opisać geometrycznie, można zinterpretować algebraicznie, i odwrotnie. Czy zatem wprowadzenie współrzędnych sprawiło, że geometria stała się niepotrzebna? A może to algebra powinna odejść do lamusa? Po co używać dwóch różnych narzędzi służących do tego samego?

W moim garażu w skrzynce z narzędziami mam młotek i duże obcęgi. Młotek służy do wbijania gwoździ w drewno, a obcęgi – do wyciągania wbitych wcześniej gwoździ. W zasadzie jednak mógłbym wbijać gwoździe obcęgami, a młotek ma z jednej strony pazur do wyciągania gwoździ. Po co mi więc oba narzędzia? Ponieważ do jednych zastosowań lepszy jest młotek, a do innych – obcęgi. Tak samo jest z geometrią i algebrą – niektóre pojęcia są bardziej naturalne na gruncie geometrii, a inne łatwiej jest opisać za pomocą algebry. Ważny jest związek łączący obie dziedziny. Gdy analiza geometryczna utknie w miejscu, możemy się przerzucić na algebrę. Gdy zaś rozważania algebraiczne zabrną w ślepy zaułek, zawsze możemy zastosować podejście geometryczne.

Geometria analityczna umożliwia wymyślanie krzywych w zupełnie nowy sposób. Wystarczy zapisać jakieś równanie i przyjrzeć się jego rozwiązaniom. Jeśli tylko nie wybraliśmy jakiegoś niemądrego równania, takiego jak $x = x$, powinniśmy dostać krzywą. (Rozwiązaniem równania $x = x$ jest cała płaszczyzna). Mogę na przykład zapisać równanie $x^3 + y^3 = 3xy$, którego rozwiązania

narysowano na rycinie 45. Tak wyglądającej krzywej, nazywanej liściem Kartezjusza, nie znajdziemy w dziełach Euklidesa. Zakres nowych krzywych, jakie można w ten sposób wymyślić, jest dosłownie nieograniczony.



Ryc. 45. Liść Kartezjusza

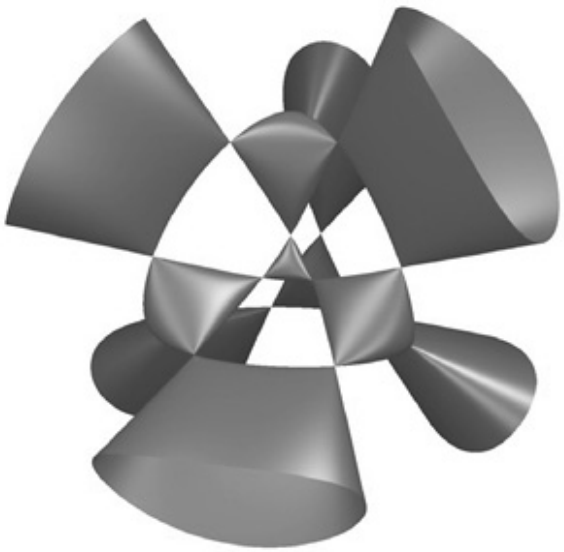
Matematycy zwykle odruchowo próbują wszystko uogólniać. Gdy któryś z nich wpadnie na jakąś interesującą koncepcję, od razu zastanawia się, czy czegoś podobnego nie dałoby się zastosować w bardziej ogólnym kontekście. Pomysł Kartezjusza daje się uogólnić czy też zmodyfikować na co najmniej trzy istotne sposoby i wszystkie one są potrzebne, żeby hipoteza Hodge'a miała jakikolwiek sens.

Najpierw zastanówmy się, co się stanie, gdy zamiast płaszczyzny użyjemy jakiejś innej przestrzeni. Trójwymiarowa przestrzeń euklidesowa ma trzy współrzędne (x, y, z) zamiast dwóch. W przestrzeni jedno równanie wyznacza zwykle powierzchnię. Dwa opisują krzywą powstałą z przecięcia odpowiadających im powierzchni. Trzy równania wyznaczają zwykle punkt. (Pisząc „zwykle", mam na myśli to, że czasami mogą się zdarzyć wyjątki, ale występują one bardzo rzadko i wymagają spełnienia specjalnych warunków. Mówiliśmy już o tego typu sytuacji na płaszczyźnie, gdy wspomnieliśmy o niemądrym równaniu $x = x$).

Również w tym wypadku zapisanie nowych równań pozwala zdefiniować nowe powierzchnie lub krzywe, których nie znajdziemy w spuściźnie Euklidesa. W XIX stuleciu było to bardzo modne zajęcie. Można było opublikować poważny artykuł naukowy poświęcony nowej powierzchni, wystarczyło tylko mieć do powiedzenia coś naprawdę interesującego na jej temat. Typowym przykładem jest powierzchnia, którą wprowadził w 1864 roku Ernst Kummer. Jej równanie wygląda następująco:

$$x^4 + y^4 + z^4 - y^2z^2 - z^2x^2 - x^2y^2 - x^2 - y^2 - z^2 + 1 = 0$$

a jej wykres pokazano na rycinie 46. Najbardziej interesującą jej cechą jest istnienie 16 „punktów podwójnych” w miejscach, w których swoim kształtem przypomina dwa stożki połączone wierzchołkami. Jest to największa możliwa liczba takich punktów dla równania czwartego stopnia i fakt ten był na tyle istotny, że zasługiwał na publikację.



Ryc. 46. Powierzchnia Kummera i jej 16 punktów podwójnych

W XIX stuleciu matematycy zakosztowali rozkoszy płynących z rozważania wielowymiarowych przestrzeni. Wcale nie musimy ograniczać się do trzech współrzędnych – możemy wprowadzić cztery, pięć, sześć… a nawet, czemu nie, milion współrzędnych. Nie są to czcze rozważania. Stoi za tym algebra dużej liczby równań i niewiadomych, a z tego typu przypadkami mamy do czynienia w matematyce bardzo często – na przykład w rozdziale 5 poświęconym hipotezie Keplera i w 8 dotyczącym mechaniki nieba. Nie jest to też zbędne uogólnienie: możliwość rozważania tego typu kwestii na gruncie geometrii i algebry jest potężnym narzędziem, którego wykorzystanie nie powinno ograniczać się do dwóch lub trzech wymiarów z tego tylko powodu, że w tych wypadkach możemy rysować wykresy i sporządzać modele.

Słowo „wymiar” brzmi być może uczenie i tajemniczo, ale w tym kontekście ma ono bardzo proste znaczenie: określa, ilu współrzędnych musimy użyć. Czterowymiarowa przestrzeń ma na przykład cztery współrzędne (x, y, z, w) i z punktu widzenia matematyki fakt ten całkowicie ją definiuje. W czterech wymiarach jedno równanie opisuje zwykle trójwymiarową „hiperpowierzchnię”, dwa równania określają powierzchnię (dwa wymiary), trzy wyznaczają krzywą (jeden wymiar), a cztery – punkt (zero wymiarów). Każde nowe równanie pozwala nam się pozbyć jednego wymiaru, czyli jednej zmiennej. Możemy więc przewidzieć, że w przestrzeni o 17 wymiarach 11 równań określa sześciowymiarowy obiekt, z wyjątkiem rzadkich (i łatwych do wykrycia) przypadków, gdy niektóre równania są nadmiarowe.

Tak zdefiniowany obiekt nazywamy rozmaitością algebraiczną. Moglibyśmy ją wprawdzie nazywać „wielowymiarową przestrzenią wyznaczoną przez układ równań algebraicznych”, ale chyba nie trzeba tłumaczyć, dlaczego nikt nie posługuje się taką nazwą.

Drugi atrakcyjny sposób uogólnienia pojęć z zakresu geometrii analitycznej polega na dopuszczeniu, by współrzędne były liczbami zespolonymi. Jak pamiętamy, w liczbach zespolonych używa się nowego rodzaju wartości oznaczanej symbolem i, która jest pierwiastkiem kwadratowym z -1. Po co wprowadzać takie komplikacje? Ponieważ w dziedzinie liczb zespolonych równania

algebraiczne stają się bardziej przewidywalne. W zbiorze liczb rzeczywistych równanie kwadratowe może mieć dwa rozwiązania albo żadnego. (Może mieć również jedno, ale w pewnym dobrze określonym sensie jest to jedno i to samo rozwiązanie występujące dwukrotnie). Natomiast w zbiorze liczb zespolonych równanie kwadratowe ma *zawsze* dwa rozwiązania (jeśli poprawnie policzymy rozwiązania wielokrotne). W wypadku niektórych zastosowań jest to bardzo pożądana cecha. Możemy powiedzieć „rozwiąż to równanie ze względu na siódmą zmienną" i mieć pewność, że takie rozwiązanie faktycznie istnieje.

Choć pod tym względem zespolona geometria algebraiczna rzeczywiście jest atrakcyjna, to ma jednak pewne cechy, do których trzeba się dopiero przyzwyczaić. W przypadku zmiennych rzeczywistych prosta może przecinać okrąg, być do niego styczna lub w ogóle nie mieć z nim punktów wspólnych, gdy mamy natomiast do czynienia ze zmiennymi zespolonymi, trzecia możliwość znika. Kiedy jednak już przywykniemy do tych zmian, okazuje się, że zespolone rozmaitości algebraiczne są wygodniejsze w użyciu od rzeczywistych. Czasami jesteśmy skazani na zmienne rzeczywiste, ale w większości wypadków lepiej się zdecydować na użycie zmiennych zespolonych. Tak czy inaczej, wiemy już teraz, czym jest zespolona rozmaitość algebraiczna.

A co z określeniem „rzutowa"? Wynika ono z trzeciego rodzaju uogólnienia i wymaga użycia nieco innego pojęcia przestrzeni. Geometria rzutowa powstała pod wpływem odkrycia perspektywy przez renesansowych malarzy i jej użycie sprawia, że proste równoległe przestają być tak wyjątkowe. W geometrii euklidesowej dwie proste albo się przecinają, albo są równoległe, czyli nie przetną się nigdy, bez względu na to, jak bardzo je przedłużymy. Wyobraźmy sobie teraz, że rozstawiliśmy sztalugi na nieskończenie dużej płaszczyźnie, w ręku trzymamy pędzel, a pudełko z farbami jest otwarte i gotowe do użycia. Przed sobą widzimy dwie proste równoległe ginące gdzieś w oddali, w blasku zachodzącego słońca, wyglądające niczym nieskończenie długie tory. Co w takiej sytuacji zobaczymy i jak to narysujemy? Na pewno nie ujrzymy dwóch

nieprzecinających się linii, ale proste, które sprawiają wrażenie, jakby się do sobie zbliżały i spotykały na horyzoncie.

Jakiej części płaszczyzny odpowiada horyzont? Tej, w której spotykają się proste równoległe. Ale przecież nie ma takiego miejsca. Horyzont jest brzegiem obrazu płaszczyzny na naszym rysunku. Jeśli świat działa jak należy, to przecież gdzieś musi być obraz brzegu płaszczyzny. No tak, ale płaszczyzna nie ma brzegu. Ciągnie się w nieskończoność. Wszystko to brzmi trochę dziwnie. Można by odnieść wrażenie, że brakuje części płaszczyzny euklidesowej. Gdy „rzutujemy" jedną płaszczyznę (tę z równoległymi torami) na drugą (płaszczyznę płótna na sztalugach), otrzymujemy na obrazie linię – horyzont – która nie jest rzutem żadnej linii istniejącej w rzeczywistości na tej płaszczyźnie.

Jest pewien sposób na poradzenie sobie z tą zagadkową anomalią – wystarczy, że na płaszczyźnie euklidesowej dodamy w nieskończoności tak zwaną prostą niewłaściwą (nazywaną też prostą w nieskończoności), odgrywającą rolę brakującego horyzontu. Wtedy wszystko stanie się znacznie prostsze. Dwie proste zawsze przecinają się w jednym punkcie – pojęcie prostych równoległych odpowiada teraz przypadkowi, gdy proste spotykają się w nieskończoności. Idea ta, jeśli się ją odpowiednio zinterpretuje, daje się przekształcić w całkowicie sensowny opis matematyczny i w efekcie uzyskujemy geometrię rzutową. Okazało się, że jest to bardzo elegancka dziedzina i matematycy w XVIII i XIX wieku wprost uwielbiali się nią zajmować. W końcu jednak powiedzieli na ten temat wszystko, co było do powiedzenia, i zainteresowanie ową dziedziną badań osłabło. Jednak w XX wieku geometria rzutowa zaczęła się znowu rozwijać, gdy matematycy uogólnili geometrię algebraiczną na wielowymiarowe przestrzenie i zaczęli wykorzystywać w ich opisie liczby zespolone. Stało się wówczas jasne, że możemy pójść na całość i zamiast ograniczać się do rozwiązań rzeczywistych w przestrzeni euklidesowej, badać rozwiązania zespolone układów równań algebraicznych w przestrzeni rzutowej.

Podsumujmy: zespolona algebraiczna rozmaitość rzutowa przypomina krzywą, którą opisuje jakieś równanie algebraiczne, z tą

różnicą, że:

- Liczba równań i zmiennych może być dowolnie duża (rozmaitość algebraiczna).
- Zmienne mogą być zespolone (zespolona).
- W pewien matematycznie sensowny sposób zmienne mogą przyjmować wartości nieskończone (rzutowa).

I tym sposobem dotarliśmy już niemal do celu, pozostaje nam do wyjaśnienia jeszcze tylko jedno określenie, z którym można szybko się rozprawić: nieosobliwa. Oznacza ono po prostu to, że rozmaitość jest gładka, pozbawiona ostrych krawędzi lub miejsc, w których jej kształt byłby bardziej skomplikowany od gładkiego kawałka przestrzeni. Powierzchnia Kummera jest na przykład osobliwa w owych 16 punktach podwójnych. Oczywiście musimy jeszcze wyjaśnić, co oznacza słowo „gładka", gdy zmienne są zespolone, a niektóre z nich mogą nawet być nieskończone, ale jest to już kwestia standardowych technik matematycznych.

Jesteśmy już niemal w połowie sformułowania hipotezy Hodge'a. Wiemy, o czym mówimy, ale nie rozumiemy jeszcze, jak zdaniem Hodge'a obiekt ten powinien się zachowywać. Musimy więc zmierzyć się teraz z najgłębszymi i najtrudniejszymi kwestiami: cyklami algebraicznymi, klasami i (szczególnie) klasami Hodge'a. Możemy jednak od razu powiedzieć, o co ogólnie tutaj chodzi. Są to narzędzia matematyczne pozwalające częściowo wyjaśnić pewną bardzo podstawową kwestię związaną z naszą uogólnioną powierzchnią, odpowiadają mianowicie na pytanie: *jakiego jest ona kształtu*? Jedyne wyrażenie, o którym jeszcze nie mówiliśmy: „wymierna kombinacja liniowa", jest właśnie poprawną odpowiedzią na tak postawione pytanie, a przynajmniej wszyscy mamy taką nadzieję.

No proszę, zobaczcie, jak daleko zaszliśmy. Rozumiemy już, czego dotyczy hipoteza Hodge'a. Wynika z niej, że mając dowolną uogólnioną powierzchnię określoną pewnymi równaniami, możemy ustalić, jakiego jest ona kształtu, dzięki wykonaniu działań

algebraicznych z wykorzystaniem czegoś, co matematycy nazywają cyklami. Mogłem to napisać już na początku tego rozdziału, ale wtedy takie stwierdzenie nie byłoby wcale bardziej zrozumiałe od ścisłego sformułowania hipotezy. Teraz, gdy już wiemy, czym jest rozmaitość, wszystko zaczyna do siebie pasować.

Zaczyna nam to przypominać również topologię. „Określanie kształtu za pomocą obliczeń algebraicznych" jest uderzająco podobne do idei Poincarégo na temat niezmienników algebraicznych przestrzeni topologicznych. W następnym kroku powinniśmy więc pomówić o topologii algebraicznej. Wśród wielu odkryć Poincarégo znajdują się również trzy ważne rodzaje niezmienników, które definiuje się za pomocą trzech pojęć: homotopia, homologia i kohomologia. Nas interesuje tutaj kohomologia i oczywiście – jakżeby mogło być inaczej – z tych trzech pojęć właśnie ona jest najtrudniejsza do wyjaśnienia.

Myślę, że musimy po prostu stawić jej czoła.

W trójwymiarowej przestrzeni o współrzędnych rzeczywistych z przecięcia sfery i płaszczyzny (przy założeniu, że się przecinają) powstaje okrąg. Sfera jest rozmaitością (będę już teraz pomijał przymiotnik „algebraiczna", gdy mowa o rozmaitościach), a także okrąg jest rozmaitością, ale okrąg zawiera się w sferze. Nazywamy go więc *podrozmaitością*. Mówiąc ogólniej, gdy do równań (wiele zmiennych, liczby zespolone, geometria rzutowa) określających pewną rozmaitość dodajemy kolejne równania, to zazwyczaj tracimy niektóre z rozwiązań – mianowicie te, które nie spełniają nowych równań. Im więcej mamy równań, tym mniejsza staje się nasza rozmaitość. Rozszerzony układ równań określa pewną część pierwotnej rozmaitości, która sama jest pełnoprawną rozmaitością – a mówiąc ściśle, podrozmaitością.

Zliczając rozwiązania równania wielomianowego, musimy czasami policzyć to samo rozwiązanie więcej niż jeden raz. Z takiego punktu widzenia zbiór rozwiązań składa się z pewnej liczby punktów i każdemu z nich możemy „przypisać" pewną liczbę – jego krotność. Możemy mieć na przykład rozwiązania 0, 1 i 2 z krotnościami, odpowiednio, 3, 7 i 4. Taki wielomian miałby postać:

$x^3(x-1)^7(x-2)^4$. Każdy z tych trzech punktów $x = 0$, 1 lub 2 jest (dość trywialną) podrozmaitością liczb zespolonych. Zatem rozwiązania tego wielomianu można opisać za pomocą listy trzech podrozmaitości, z przyczepionymi do nich, niczym etykiety, liczbami naturalnymi.

Cykl algebraiczny jest czymś podobnym. Zamiast pojedynczych punktów posługujemy się tym razem dowolnymi skończonymi listami podrozmaitości. Do każdej z nich możemy doczepić etykietę z liczbą, która nie musi wcale być liczbą całkowitą dodatnią. Może to być liczba całkowita ujemna, wymierna, rzeczywista, a nawet zespolona. Z różnych względów w hipotezie Hodge'a w roli etykiet występują liczby wymierne. Właśnie do tego odnosi się określenie „wymierna kombinacja liniowa". I tak, gdyby naszą wyjściową rozmaitością była na przykład sfera jednostkowa w jedenastowymiarowej przestrzeni, to taka lista mogłaby wyglądać następująco:

- siedmiowymiarowa hipersfera (o takich i takich równaniach) z etykietą 22/7,
- torus (o takich i takich równaniach) z etykietą –4/5,
- krzywa (o takim i takim równaniu) z etykietą 413/6.

Nie próbujcie sobie tego wyobrazić, a jeśli już musicie, spójrzcie na to oczami autora rysunków komiksowych, rysując w myśli schematycznie trzy nieregularne plamy z maleńkimi etykietami. Każdy kadr takiego komiksu, czyli każda lista, stanowi jeden cykl algebraiczny.

Po co się tak męczyć i wymyślać coś tak abstrakcyjnego? Ponieważ w ten sposób możemy uchwycić kluczowe aspekty wyjściowej rozmaitości algebraicznej. Specjaliści od geometrii algebraicznej zastosowali w tym celu pewną sztuczkę podpatrzoną u topologów.

W rozdziale 10 poświęconym hipotezie Poincarégo mówiliśmy o mrówce, dla której całym wszechświatem jest pewna powierzchnia. W jaki sposób mogłaby ona ustalić, jaki kształt ma jej

wszechświat, skoro nie może wyjść poza niego i popatrzeć z boku? W szczególności, jak mogłaby odróżnić sferę od torusa? Przedstawione wówczas rozwiązanie opierało się na wykorzystaniu pętli – topologicznych tras autobusowych. Mrówka może przesuwać takie pętle w różne miejsca, sprawdzać, co się stanie, gdy je ze sobą połączy, i wyliczać algebraiczny niezmiennik przestrzeni, który nazywamy jej grupą podstawową. Słowo „niezmiennik" oznacza, że przestrzenie równoważne topologicznie muszą mieć taką samą grupę podstawową. Jeśli grupy te są różne, to mamy do czynienia z różnymi przestrzeniami. Właśnie taki niezmiennik naprowadził Poincarégo na sformułowanie jego hipotezy. Jednak wymaganie od biednej mrówki, żeby zbadała wszystkie możliwe pętle w swym wszechświecie, może być zadaniem ponad jej siły – wynika to stąd, że wyznaczenie grupy podstawowej wymaga uporania się z naprawdę skomplikowanymi problemami matematycznymi. Poincaré badał również pewien bardziej praktyczny niezmiennik. Przesuwanie pętli w różne miejsca nazywamy homotopią. Podejście alternatywne ma podobną nazwę: homologia.

Omówimy tu najprostszą, najbardziej konkretną wersję homologii. Topolodzy szybko usprawnili tę wersję, poprawili co trzeba, uogólnili i przekształcili w potężną machinę matematyczną zwaną algebrą homologiczną. Ta najprostsza wersja pozwala jedynie wyrobić sobie ogólne pojęcie na temat tej dziedziny, ale nam to w zupełności wystarczy.

Nasza mrówka zapragnęła sporządzić mapę swojego wszechświata, postanawia więc go gruntownie zbadać. Korzystając z tej samej metody co kartografowie, pokrywa cały wszechświat siecią trójkątów. Musi przy tym jednak spełnić jeden kluczowy warunek: żaden z trójkątów nie może obejmować otworu w powierzchni. Aby mieć pewność, że tak jest, mrówka tworzy kolejne trójkąty, rozkładając na powierzchni gumowe łatki, jak podczas naprawiania przebitej dętki. Dzięki temu każdy trójkąt ma dobrze określone wnętrze, które z topologicznego punktu widzenia jest takie samo jak wnętrze zwyczajnego trójkąta na płaszczyźnie. Topolodzy nazywają takie łatki dyskami topologicznymi, ponieważ są one tożsame z kołem, czyli okręgiem wraz z jego wnętrzem. Aby zrozumieć,

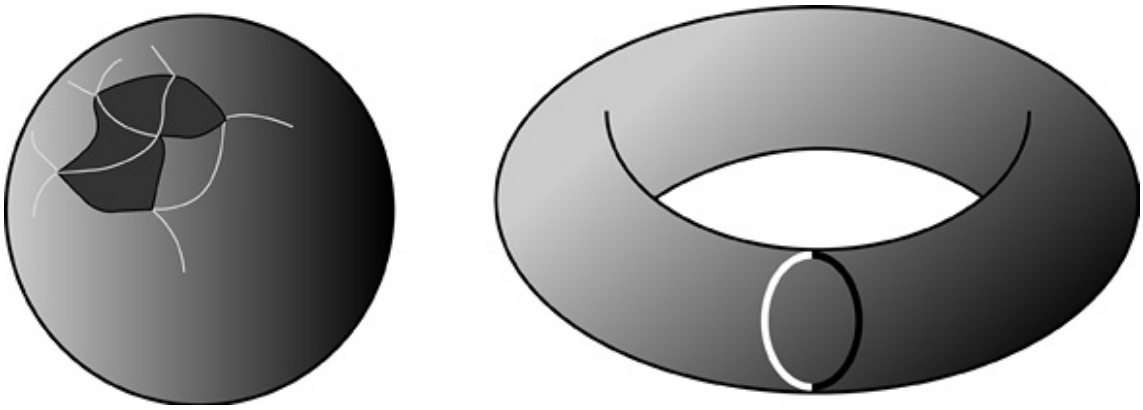
dlaczego tak jest, wystarczy popatrzeć na rycinę 36, na której pokazano, jak w sposób ciągły można przekształcić trójkąt w koło. Tego rodzaju łatki nie można rozciągnąć na trójkącie otaczającym jakiś otwór, ponieważ taka dziura tworzy tunel łączący zewnętrze trójkąta z jego wnętrzem. W takiej sytuacji łatka musiałaby wychodzić poza powierzchnię, a mrówce nie wolno poza nią wyjść.

Tym sposobem mrówka dokonała *triangulacji* całego swojego wszechświata. Dzięki wprowadzeniu wymogu użycia łatek mamy pewność, że lista wszystkich trójkątów wraz z informacją o tym, które z nich się ze sobą stykają, pozwoli nam odtworzyć topologię powierzchni, czyli jej kształt w sensie równoważności topologicznej. Gdybyśmy wybrali się do Ikei i kupili płasko spakowany Wszechświat Mrówki zawierający odpowiednio oznaczone trójkąty, a następnie skleili krawędź A z AA, B z BB i tak dalej, to moglibyśmy skonstruować całą tę powierzchnię. Mrówka nie może wyjść poza powierzchnię, nie sporządzi więc takiego modelu, ale może zrobić wszystko, żeby przynajmniej w zasadzie jej mapa zawierała całą potrzebną do tego informację. Aby ją uzyskać, musi przeprowadzić obliczenia. Nie oznacza to już konieczności rozważania nieskończonej liczby wszystkich możliwych pętli, ale i tak ma nad czym rozmyślać – musi rozważyć wszystkie pętle przebiegające przez krawędzie sieci, którą skonstruowała.

W przypadku homotopii interesuje nas, czy daną pętlę da się skurczyć w sposób ciągły do rozmiaru punktu. Analizując homologię, stawiamy inne pytanie: czy dana pętla jest brzegiem topologicznego dysku? Innymi słowy, czy można ułożyć jedną lub więcej trójkątnych łatek w taki sposób, żeby powstał obszar bez otworów, a granicą tego obszaru była rozważana pętla?

Na rycinie 47 (po lewej) pokazano fragment triangulacji sfery, jedną z pętli i topologiczny dysk, którego brzeg odpowiada tej pętli. Stosując odpowiednią technikę, można udowodnić, że *każda* pętla w triangulacji sfery tworzy brzeg – trójkątne łatki, czyli mówiąc ogólniej, topologiczne dyski, są wykrywaczami otworów, a przecież intuicyjnie wiemy, że sfera nie ma dziur. Torus natomiast ma jeden otwór i faktycznie, niektóre pętle na torusie nie tworzą brzegu. Na tej samej rycinie (po prawej stronie) pokazano przykład takiej pętli

przechodzącej przez otwór w środku torusa. Możemy zatem powiedzieć, że przeglądając listę pętli i sprawdzając, które z nich tworzą brzeg, mrówka może odróżnić świat sferyczny od toroidalnego.



Ryc. 47. Po prawej: Fragment triangulacji sfery, jedna z pętli (zaznaczona czarną linią) i dysk, którego brzegiem jest zaznaczona pętla (obszar zacieniony). Po lewej: Pętla na torusie, która nie jest brzegiem żadnego dysku (biały fragment znajduje się z tyłu).

Gdyby nasza mrówka była tak sprytna jak Poincaré i inni współcześni mu topolodzy, to mogłaby przekształcić tę ideę w elegancki niezmiennik topologiczny – w grupę homologii swojej powierzchni. Podstawowa idea polega na „dodawaniu" dwóch pętli przez narysowanie ich obydwu. Jednak w ten sposób powstanie coś, co nie jest pętlą, musimy więc wrócić do początku i zacząć wszystko od nowa. Tak naprawdę musimy cofnąć się daleko w przeszłość, do czasów, gdy dopiero poznawaliśmy algebrę. Gdy rozpoczynałem naukę matematyki, na jednej z pierwszych lekcji nauczyciel powiedział nam, że możemy dodawać jabłka do jabłek i uzyskać całkowitą ich liczbę, ale nie można dodawać jabłek do gruszek, no chyba że interesuje nas ogólna liczba owoców.

Tak to wygląda w arytmetyce, choć nawet tam musimy uważać, żeby nie użyć tego samego jabłka dwukrotnie, natomiast w algebrze takie ograniczenia nie obowiązują. Możemy dodawać jabłka do gruszek, nie mieszając ich jednak ze sobą. Mówiąc szczerze, w zawansowanej matematyce na porządku dziennym dodaje się do siebie rzeczy, których, zdawałoby się, nikt przy zdrowych zmysłach nie może wymyślić, nie mówiąc już o ich dodawaniu. Możliwość przeprowadzania tego typu operacji okazuje się niezwykle przydatna i ważna, a postępujący tak matematycy nie są wcale szaleni – przynajmniej nie pod tym względem.

Aby zrozumieć koncepcje, które łączy ze sobą hipoteza Hodge'a, musimy móc dodawać jabłka do gruszek bez konieczności pakowania ich do jednego pudełka z napisem „owoce". Okazuje się, że nie jest to wcale takie trudne. Jedyny kłopot polega na przyznaniu, że jest nam do czegoś potrzebne. Wielu z nas spotkało się już wcześniej z pewną odmianą takiego podejścia. Na wspomnianej lekcji dowiedzieliśmy się też, że w matematyce nieznane wartości oznacza się literami, a różne litery odpowiadają różnym niewiadomym. Jeśli mamy *a* jabłek, a potem dostaniemy jeszcze następne *a* jabłek, to w sumie będziemy ich mieli $a + a = 2a$. Taka metoda działa dla dowolnej liczby jabłek. Jeśli do $3a$ jabłek dodamy ich jeszcze $2a$, to będziemy mieli $5a$ jabłek i będzie to prawdą bez względu na to, ile faktycznie liczba ta wynosi. Nie ma tu również znaczenia wybór symboli ani to, co one oznaczają: jeśli do $3b$ gruszek dodamy ich jeszcze $2b$, to uzyskamy wynik $5b$. A co w sytuacji, gdy mamy $3a$ jabłek i $2b$ gruszek? Ile to jest $3a$ plus $2b$?

$$3a + 2b.$$

I już. Nie można uprościć takiej sumy, tak by uzyskać 5 czegoś tam – a przynajmniej nie można tego zrobić bez przekształceń obejmujących wprowadzenie nowej kategorii (owoce) i nowych równań. To jest najlepsze, co możemy osiągnąć, i musimy się z tym pogodzić. Jednak po zrobieniu tego kroku możemy wyliczać sumy takie jak:

$$(3a+2b)+(5a-b)=8a+b$$

bez konieczności wprowadzania dodatkowych pojęć czy nowych rodzajów owoców.

Należy tu jednak pamiętać o pewnych zastrzeżeniach. Wspomnieliśmy już o tym, że dodając jedno jabłko do drugiego, otrzymamy dwa jabłka tylko wtedy, gdy to drugie nie było tym samym co pierwsze. To samo zastrzeżenie obowiązuje dla bardziej skomplikowanych kombinacji jabłek i gruszek. W algebrze zakłada się w obliczeniach, że wszystkie jabłka są różne. Prawdę mówiąc, w większości wypadków jest to sensowne założenie, nawet wtedy, gdy oba jabłka – czy cokolwiek, co do siebie dodajemy – mogą być w istocie jednym i tym samym jabłkiem. Jedno jabłko plus to samo jabłko równa się jabłko z krotnością dwa.

Gdy się już przyzwyczaimy do takiego podejścia, możemy je stosować w dowolnej sytuacji. Świnia plus ta sama świnia równa się świnia z krotnością dwa: świnia + świnia = 2 świnia, bez względu na to, czym jest owa świnia. Świnia plus krowa równa się świnia + krowa. Trójkąt plus trzy okręgi równa się trójkąt + 3 okrąg. Superhipersfera plus trzy hipereliptyczne kwazistogi równa się

superhipersfera + 3 hipereliptycznykwazistóg

cokolwiek dana terminologia oznacza (w naszym przykładzie: zupełnie nic).

Możemy nawet dopuścić stosowanie liczb ujemnych i używać określeń takich jak trzy świnie minus jedenaście krów: 3 świnia – 11 krowa. Nie mam pojęcia, jak wygląda minus jedenaście krów, ale jestem pewien, że jeśli dodam do nich jeszcze sześć krów, to otrzymam minus pięć krów[88]. Jest to sformalizowana gra, w którą gra się przy użyciu symboli, i nie jest do tego potrzebna żadna bardziej realistyczna interpretacja, ba, nie byłaby ona użyteczna, a często nawet w ogóle możliwa. Możemy dopuścić użycie liczb rzeczywistych: π świń minus $\sqrt{2}$ krów. Liczb zespolonych. Dowolnego

innego wyrafinowanego rodzaju liczb, jakie matematycy wymyślili lub wymyślą w przyszłości. Podejście to może nabrać trochę większej wartości, jeśli przyjmiemy, że liczby są *etykietkami* doczepionymi do świń i krów. W takim razie π świń minus $\sqrt{2}$ krów można sobie wyobrazić jako świnię z etykietką π stojącą obok krowy z nalepką $-\sqrt{2}$. Reguły arytmetyczne stosujemy do etykietek, nie do zwierząt.

Hipoteza Hodge'a wykorzystuje właśnie tego rodzaju konstrukcję, z dodatkowymi bajerami. Zamiast zwierząt występują w niej krzywe, powierzchnie i ich wielowymiarowe odpowiedniki. Choć może się to wydawać dziwne, nie wynika z tego jakiś abstrakcyjny nonsens, ale głęboki związek między topologią, algebrą, geometrią i analizą matematyczną.

Aby przygotować formalizm odpowiedni dla homologii, musimy w jakiś sposób dodawać do siebie pętle, ale nie tak, jak robiliśmy to w przypadku grupy podstawowej. Teraz trzeba je dodawać tak, jak nauczono mnie tego w szkole. Powinniśmy po prostu zapisać pętle i postawić między nimi znak „+". Aby to miało sens, musimy się zająć nie pojedynczymi pętlami, ale całymi ich skończonymi zbiorami. Do każdej pętli doczepmy jeszcze etykietkę z liczbą całkowitą mówiącą o tym, jak często dana pętla występuje. Nazwijmy taki zbiór z etykietkami *cyklem*. Teraz mrówka może dodać do siebie dowolne dwa cykle, grupując ze sobą odpowiednie elementy i dodając do siebie ich etykiety – otrzyma w ten sposób jakiś inny cykl. Może w związku z tym w rozdziale 10 powinniśmy byli wyobrazić sobie, że mrówka jeździ na bicyklu, a nie autobusem?

Gdy konstruowaliśmy grupę podstawową, stosując „dodawanie" rozumiane jako łączenie ze sobą pętli, natrafiliśmy na pewien szkopuł. Dodanie pętli trywialnej do innej pętli nie prowadziło do uzyskania *dokładnie* tej samej pętli, pętla zerowa więc zachowywała się niezgodnie z oczekiwaniami. Wynik dodania pętli do pętli przeciwnej nie był tym samym co pętla trywialna, zatem elementy odwrotne również nie działały poprawnie. Poradziliśmy sobie z tym problemem, przyjąwszy, że dwie pętle są takie same, jeśli jedną z nich można przekształcić w drugą.

W przypadku homologii problem taki nie występuje. Istnieje cykl zerowy (wszystkie etykietki mają wartość zero) i każdy cykl ma cykl przeciwny (wystarczy pozamieniać liczby na wszystkich etykietach na liczby ze znakiem przeciwnym), otrzymujemy więc w ten sposób grupę. Kłopot jednak w tym, że jest to nieodpowiednia grupa. Nie mówi nam ona nic o topologii przestrzeni. Aby sobie z tym poradzić, musimy zastosować inną podobną sztuczkę i rozluźnić nieco definicję tego, jaki cykl uważamy za zerowy. Mrówka podzieliła przestrzeń na trójkątne obszary, których brzeg jest z topologicznego punktu widzenia dość trywialny: przesuwając go w kierunku środka obszaru, możemy skurczyć taki brzeg do rozmiaru punktu. Wprowadźmy więc wymaganie, że takie cykle brzegowe należy uważać za równoważne z cyklem zerowym. Przypomina to trochę zmianę znaczenia liczb naturalnych w arytmetyce modularnej, gdy udajemy, że liczba 12 jest nieistotna i w związku z tym można ją zamienić na zero. Tutaj zmieniamy znaczenie cykli w homologii, udając, że każdy cykl brzegowy jest nieistotny.

Z tego założenia wynikają daleko idące konsekwencje. Za jego sprawą algebra cykli staje się wrażliwa na topologię przestrzeni. Grupa cykli z cyklem brzegowym w roli modułu jest użytecznym niezmiennikiem topologicznym, grupą homologii powierzchni. Na pierwszy rzut oka mogłoby się wydawać, że zależy ona od wyboru triangulacji, ale tak jak w przypadku charakterystyki Eulera, różne triangulacje tej samej powierzchni prowadzą do uzyskania tej samej grupy homologii. W ten sposób mrówka wymyśliła niezmiennik algebraiczny pozwalający jej odróżnić od siebie różne powierzchnie. Jest on trochę nieporęczny, ale nie ma dobrego niezmiennika, który nie wymagałby przeprowadzenia na jakimś etapie trudnych obliczeń. Ten jest tak skuteczny, że pozwala na odróżnienie nie tylko sfery od torusa, ale także torusa z dwoma otworami od takiego, który ma pięć otworów – a także od torusa o dowolnej innej liczbie dziur.

Homologia może się wydawać dość skomplikowana, ale dzięki niej powstał bogaty zbiór topologicznych niezmienników, a sama opiera się na prostych geometrycznych pojęciach: pętle, brzegi, grupowanie

zbiorów i wykonywanie operacji arytmetycznych na etykietach. Wziąwszy pod uwagę, że biedna mrówka jest uwięziona na swojej powierzchni, musimy przyznać, że to zdumiewające, iż może w ogóle powiedzieć coś ciekawego na temat kształtu swojego wszechświata, rozciągając jedynie trójkątne łatki, sporządzając mapy i wykonując obliczenia algebraiczne.

Istnieje naturalny sposób rozszerzenia homologii na wyższe wymiary. Trójwymiarowym odpowiednikiem trójkąta jest czworościan – jest to bryła o czterech wierzchołkach, sześciu krawędziach i czterech trójkątnych ścianach. Jego wnętrzem jest pojedyncza trójwymiarowa „ściana". Mówiąc ogólniej, w *n* wymiarach możemy zdefiniować *n*-wymiarowy sympleks o *n* + 1 wierzchołkach, które połączone parami tworzą wszystkie możliwe krawędzie – one z kolei tworzą trójkąty, które po połączeniu dają czworościany, i tak dalej. Możemy teraz bez trudu zdefiniować cykle, brzegi i homologie, a także zestawić grupę, dodając do siebie cykle (a właściwie ich klasy homologii). Otrzymujemy w istocie cały ciąg grup: jedną dla cykli zerowymiarowych (punktów), jedną dla cykli jednowymiarowych (odcinków), jedną dla cykli dwuwymiarowych (trójkątów) i tak dalej, aż do wymiaru samej przestrzeni. Tak uzyskujemy zerową, pierwszą, drugą i tak dalej grupę homologii danej przestrzeni. Mówiąc w uproszczeniu, grupy te pozwalają precyzyjnie zdefiniować pojęcie otworów o różnej liczbie wymiarów i ustalić, czy one w ogóle istnieją, ile ich jest i jak są ze sobą związane.

Tak zatem wygląda homologia i jest ona już niemal tym, co jest nam potrzebne do zrozumienia *sformułowania* hipotezy Hodge'a. Tak naprawdę interesuje nas jednak pojęcie blisko związane z homologią, a mianowicie: *kohomologia*. W 1893 roku Poincaré zwrócił uwagę na dziwny zbieg okoliczności w homologii dowolnej rozmaitości. Otóż okazało się, że lista grup homologii wygląda tak samo, gdy odczytuje się ją od końca. Dla rozmaitości o, powiedzmy, pięciu wymiarach zerowa grupa homologii jest taka sama jak piąta, pierwsza wygląda jak czwarta, a druga – jak trzecia. Doszedł do wniosku, że nie może to być jedynie kwestia przypadku, i wyjaśnił to zjawisko za pomocą triangulacji dualnej – pojęcia, z którym

spotkaliśmy się już w rozdziale 4 przy okazji omawiania map. Chodzi tu o drugą triangulację, która powstaje dzięki zastąpieniu każdego trójkąta wierzchołkiem, każdej krawędzi między dwoma trójkątami krawędzią łączącą odpowiadające im nowe wierzchołki i każdego wierzchołka trójkątem – analogicznie do konstrukcji pokazanej na rycinie 10. Zwróćmy uwagę na to, że teraz wymiary pojawiają się w odwrotnej kolejności: dwuwymiarowe trójkąty stają się zerowymiarowymi punktami i odwrotnie, natomiast jednowymiarowe krawędzie pozostają jednowymiarowymi krawędziami, ponieważ są w środku.

Okazuje się, że dobrze jest odróżniać od siebie te dwie listy, mimo iż prowadzą do uzyskania tych samych niezmienników. Po uogólnieniu całej konstrukcji i wyrażeniu jej za pomocą pojęć abstrakcyjnych triangulacje znikają i mówienie o triangulacji dualnej traci sens. Pozostają tylko dwa ciągi niezmienników topologicznych, które nazywamy grupami homologii i grupami kohomologii. Każde pojęcie w homologii ma swój dualny odpowiednik, który zwykle nazywa się tak samo, tylko z dodanym „ko-” na początku. Zatem zamiast cykli mamy kocykle, a dwóm cyklom homologicznym odpowiadają dwa kocykle kohomologiczne. Klasy, o których mowa w hipotezie Hodge’a, to klasy kohomologii. Są one zbiorami kocykli, które są kohomologiczne względem siebie.

Homologia i kohomologia nie mówią nam o kształcie przestrzeni topologicznej wszystkiego, co chcielibyśmy wiedzieć – różne przestrzenie mogą mieć tę samą homologię i kohomologię – ale dostarczają wielu cennych informacji, a poza tym dysponujemy odpowiednim formalizmem pozwalającym je obliczyć i wykorzystać.

Rozmaitość algebraiczna, bez względu na to, czy jest rzeczywista, zespolona czy rzutowa, jest przestrzenią topologiczną, a zatem ma jakiś kształt. Aby dowiedzieć się czegoś użytecznego na temat tego kształtu, musimy myśleć jak topolodzy i wyznaczyć grupy homologii i kohomologii. Jednak naturalnymi składnikami geometrii algebraicznej wcale nie są obiekty geometryczne, takie jak triangulacje i cykle. Składnikami tymi są obiekty, które najłatwiej

można opisać za pomocą równań algebraicznych. Spójrzmy jeszcze raz na równanie powierzchni Kummera. Jak można by je powiązać z triangulacją? We wzorze tej powierzchni nie ma nic, co mogłoby się kojarzyć z trójkątami.

Może powinniśmy zacząć jeszcze raz od początku i zamiast trójkątów posłużyć się naturalnymi elementami składowymi rozmaitości, czyli podrozmaitościami, które definiuje się, dodając dodatkowe równania. Musimy zatem przyjąć nową definicję cykli: zamiast zbiorów trójkątów z etykietami z liczbami całkowitymi użyjemy teraz podrozmaitości z dowolnymi etykietami, które odegrają swoją rolę. Z różnych powodów – a głównie dlatego, że hipoteza Hodge'a jest nieprawdziwa, gdy użyje się etykiet z liczbami całkowitymi – najlepszym wyborem są liczby wymierne. Pytanie Hodge'a sprowadza się do następującej kwestii: czy taka nowa definicja homologii i kohomologii obejmuje wszystko, co zawiera się w topologicznej definicji tych pojęć? Jeśli hipoteza ta jest prawdziwa, to algebraiczne narzędzie wykorzystujące pojęcie cyklu jest tak ostre, że może się równać z topologicznym mieczem kohomologii. Jeśli nie jest spełniona, to algebraiczny cykl jest tępym, nieprzydatnym do niczego przyrządem.

Tyle tylko, że… och, przepraszam, ale trochę się zagalopowałem. Hipoteza Hodge'a głosi, że wystarczy posłużyć się konkretnym *rodzajem* cyklu algebraicznego, który istnieje w ramach klasy Hodge'a. Aby to wyjaśnić, do naszego i tak już bogatego zasobu wiedzy musimy dodać jeszcze jeden element: analizę matematyczną. Jednym z najważniejszych pojęć analizy matematycznej jest równanie różniczkowe, opisujące szybkość zmiany zmiennych (zob. rozdz. 8). Niemal wszystkie modele fizyki matematycznej XVIII, XIX i XX wieku opisują świat za pomocą równań różniczkowych – ba, nawet większość modeli powstających w XXI wieku opiera się na tym samym podejściu. Hodge badał tego typu równania w latach trzydziestych XX wieku i opracował zupełnie nowy zestaw metod, który nazywamy obecnie teorią Hodge'a. Łączy się ona w naturalny sposób z licznymi innymi metodami analizy matematycznej i topologii.

Pomysł Hodge'a polegał na użyciu równania różniczkowego do podziału klas kohomologii na różne typy. Każdy element ma dodatkową strukturę, która przydaje się w analizie problemów topologicznych. Elementy te definiuje się za pomocą równania różniczkowego, które pojawiło się pod koniec XVIII wieku głównie w pracach Pierre'a Simona de Laplace'a. Laplace zajmował się głównie mechaniką nieba, czyli badaniem ruchu i kształtu planet, księżyców, komet oraz gwiazd. W 1783 roku pracował nad szczegółowym opisem kształtu Ziemi. Wiedziano już wówczas, że Ziemia nie jest okrągła, lecz spłaszczona na biegunach i ma kształt sferoidy – przypomina piłkę plażową, na której ktoś usiadł. Jednak nawet taki opis nie oddaje niektórych szczegółów. Laplace opracował metodę wyznaczania jej kształtu z dowolną wymaganą dokładnością w oparciu o pewną wielkość fizyczną odpowiadającą ziemskiemu polu grawitacyjnemu – nie chodzi tu o samo pole, ale jego potencjał. Potencjał grawitacyjny jest miarą energii zawartej w grawitacji, wielkością liczbową zdefiniowaną w każdym punkcie przestrzeni. Siła grawitacji zawsze działa w takim kierunku, w którym potencjał maleje najszybciej, a wielkość tej siły jest równa szybkości zmniejszania się potencjału.

Potencjał spełnia równanie Laplace'a. Ogólnie rzecz biorąc, oznacza to, że w razie nieobecności materii – to znaczy w próżni – średnia wartość potencjału w obszarze bardzo małej kuli jest równa jego wartości w środku kuli. Mamy tu do czynienia ze swego rodzaju demokracją: nasza wartość jest średnią wartości naszych sąsiadów. Każde rozwiązanie równania Laplace'a nazywamy funkcją harmoniczną. Specjalne rodzaje klas kohomologii Hodge'a to te, które łączy szczególny związek z funkcjami harmonicznymi. Teoria Hodge'a, związana z badaniem tych rodzajów, zapoczątkowała rozwój głębokiego i wspaniałego działu matematyki: związków między topologią przestrzeni a pewnymi szczególnymi równaniami różniczkowymi określonymi na tej przestrzeni.

Teraz mamy już wszystkie składniki. Z naszych rozważań wynika, że hipoteza Hodge'a mówi o istnieniu silnego związku między trzema filarami współczesnej matematyki: algebrą, topologią i analizą matematyczną. Weźmy dowolną rozmaitość. Aby zrozumieć jej

kształt (a więc mówimy o topologii i klasach kohomologii), musimy wybrać szczególne przypadki jej klas kohomologii (a do tego potrzebna jest analiza matematyczna, pozwalająca na zdefiniowanie klas Hodge'a przy użyciu równań różniczkowych). Te szczególne rodzaje klas kohomologii można znaleźć z wykorzystaniem podrozmaitości (teraz jesteśmy na gruncie algebry i mówimy o dorzuceniu kilku dodatkowych równań i szukaniu cykli algebraicznych). Innymi słowy, aby odpowiedzieć na nasze pytanie topologiczne o kształt rozmaitości, musimy je przekształcić w pytanie z zakresu analizy matematycznej, a następnie znaleźć odpowiedź, wykorzystując do tego algebrę.

Dlaczego jest to takie ważne? Hipoteza Hodge'a jest propozycją wzbogacenia geometrii o dwa nowe narzędzia: niezmienniki topologiczne i równanie Laplace'a. Nie jest to tak naprawdę hipoteza mówiąca o jakimś twierdzeniu matematycznym, ale stwierdzenie odnoszące się do nowego rodzaju narzędzi. Jeśli dowiedziemy jego prawdziwości, narzędzia te automatycznie zyskają nowe znaczenie i być może uda się je wykorzystać do znalezienia odpowiedzi na ogromną liczbę pytań. Oczywiście, może się okazać, że hipoteza ta jest nieprawdziwa. Byłoby to rozczarowujące, ale lepiej jest mieć świadomość, jakie są ograniczenia stosowanych narzędzi, niż raz za razem uderzać się nimi boleśnie w palec.

Skoro mamy już pojęcie, jaka jest natura hipotezy Hodge'a, możemy rozejrzeć się za potwierdzającymi ją dowodami. Co wiemy na ten temat? Bardzo niewiele.

W 1924 roku, zanim Hodge wysunął swoją hipotezę, Salomon Lefschetz udowodnił twierdzenie, które sprowadza się do hipotezy Hodge'a dla dwuwymiarowej kohomologii dowolnej rozmaitości. Po przeprowadzeniu standardowych obliczeń z zakresu topologii algebraicznej można na tej podstawie udowodnić, że hipoteza - Hodge'a jest spełniona dla rozmaitości jedno-, dwu- i trójwymiarowych. Dla rozmaitości o większej liczbie wymiarów znamy tylko kilka szczególnych przypadków zgodnych z hipotezą Hodge'a.

Hodge sformułował pierwotnie swoją hipotezę z wykorzystaniem etykiet z liczbami całkowitymi. W 1961 roku Michael Atiyah i Friedrich Hirzebruch udowodnili, że w wyższych wymiarach taka postać hipotezy jest nieprawdziwa. Obecnie zatem hipotezę Hodge'a interpretuje się z wykorzystaniem etykiet z liczbami wymiernymi. Dla takiej wersji istnieje pewna liczba obiecująco wyglądających dowodów. Najsilniejszym dowodem przemawiającym za poprawnością tej hipotezy jest fakt, że udało się udowodnić jeszcze bardziej skomplikowane twierdzenie, znane jako „algebraiczność miejsc geometrycznych Hodge'a", które jest najważniejszą wynikającą z niej konsekwencją. Dowód ten przedstawili w 1995 roku Eduardo Cattani, Pierre Deligne i Arnoldo Kaplan *bez* zakładania poprawności hipotezy Hodge'a.

Warto jeszcze wspomnieć o pewnej atrakcyjnej hipotezie z zakresu teorii liczb, która jest odpowiednikiem hipotezy Hodge'a. Chodzi o hipotezę Tate'a – nazwaną tak od nazwiska Johna Tate'a – łączącą geometrię algebraiczną z teorią Galois, czyli z pojęciami, które pozwoliły udowodnić, że nie ma żadnego wzoru algebraicznego umożliwiającego rozwiązanie równania wielomianowego stopnia 5. Sformułowanie tej hipotezy jest dość skomplikowane i odwołuje się do jeszcze innego rodzaju kohomologii. Z pewnych względów, niezależnych od hipotezy Hodge'a, możemy mieć nadzieję, że hipoteza Tate'a okaże się prawdziwa, ale do tej pory nie zostało to jeszcze udowodnione. Mamy jednak przynajmniej jakąś inną sensowną hipotezę związaną z hipotezą Hodge'a, nawet jeśli na razie wydaje się równie trudna od udowodnienia.

Hipoteza Hodge'a jest jednym z tych denerwujących matematycznych stwierdzeń, dla których nie mamy wielu dowodów przemawiających za ich poprawnością lub niepoprawnością, a na domiar złego dowody te nie są szczególnie przekonujące. Niewątpliwie istnieje niebezpieczeństwo, że hipoteza ta może się okazać nieprawdziwa. Być może jest jakaś rozmaitość o milionie wymiarów dowodząca niepoprawności hipotezy Hodge'a z powodów, które sprowadzają się do długiego ciągu tak zawiłych obliczeń, że nigdy nie uda się ich wykonać. Gdyby tak było, hipoteza Hodge'a mogłaby być nieprawdziwa w zasadzie z bardzo prozaicznych

powodów, ale udowodnienie tego byłoby praktycznie niemożliwe. Znam kilku specjalistów od geometrii algebraicznej, którzy podejrzewają, że tak jest. Jeśli to prawda, to w dającej się przewidzieć przyszłości ów milion dolarów będzie bezpiecznie leżał w banku.

88 W tym miejscu przypomina mi się stary matematyczny dowcip. Biolog, statystyk i matematyk siedzą sobie przy kawiarnianym stoliku i przyglądają się toczącemu się wokół nich życiu. Jakiś mężczyzna wchodzi z kobietą do budynku po drugiej stronie ulicy. Po dziesięciu minutach oboje wychodzą w towarzystwie dziecka.

– O, rozmnożyli się – zauważa biolog.

– Nie – oponuje statystyk – to błąd obserwacyjny. Średnio rzecz biorąc, do budynku weszły i z niego wyszły dwie i pół osoby.

– Ależ nie – protestuje matematyk – przecież to oczywiste. Jeśli teraz ktoś wejdzie do środka, to budynek będzie pusty.

16. Co dalej?

Przewidywanie jest bardzo trudne, zwłaszcza gdy dotyczy przyszłości[89] – jako autora tego celnego aforyzmu wskazuje się laureata Nagrody Nobla, fizyka Nielsa Bohra, lub zawodnika i trenera baseballu amerykańskiego Yogiego Berrę[90]. Nie należy jednak zapominać, że sam Yogi Berra przyznał: „Nigdy nie powiedziałem większości tego, co powiedziałem". Rzekomo. Arthur C. Clarke, znany autor książek fantastycznonaukowych i scenariusza do filmu *2001: Odyseja kosmiczna*, a także dalszych części tego cyklu, był również futurologiem – pisał książki, w których starał się przewidzieć rozwój techniki i społeczeństwa. W wydanej w 1962 roku książce *Profiles of the Future* (*Charakterystyka przyszłości*) wśród licznych przewidywań pojawiły się i takie:

- do roku 1970 – zrozumienie języka wielorybów i delfinów,
- do roku 1990 – ujarzmienie mocy syntezy jądrowej,
- do roku 1990 – wykrycie fal grawitacyjnych,
- do roku 2000 – kolonizacja planet.

Nic z tego jeszcze się nie spełniło. Równocześnie jednak niektóre z jego przewidywań okazały się trafione:

- do roku 1980 – lądowanie na planetach (choć niewykluczone, że miał na myśli dotarcie ludzi na inne planety),
- do roku 1970 – urządzenia do automatycznego tłumaczenia (może jeszcze jest za wcześnie na ogłaszanie ich powstania, ale Google ma tego typu możliwości),
- do roku 1990 – osobiste radio (telefony komórkowe działają na tej zasadzie).

Przewidział również, że do 2000 roku będziemy mieli globalną bibliotekę i być może jesteśmy bliżej osiągnięcia tego celu, niż sądziliśmy jeszcze kilka lat temu, ponieważ jest to jedna z wielu funkcji Internetu. Powstanie chmur obliczeniowych może być pierwszym krokiem na drodze do budowy jednego gigantycznego komputera, z którego będą korzystali wszyscy mieszkańcy planety. Clarke nie dostrzegł kilku bardzo ważnych procesów, takich jak konstrukcja komputerów i opanowanie inżynierii genetycznej, choć przewidział, że ludzkość osiągnie te umiejętności do 2030 roku. Biorąc pod uwagę, że nawet Clarke'owi udawało się przewidywać przyszłość jedynie ze zmiennym szczęściem, musimy przyznać, iż próba ustalenia, jaki będzie los wielkich problemów matematycznych, wydaje się dość zuchwała. Jednak bazując na posiadanych informacjach, możemy spróbować zgadnąć, jak potoczą się dalsze wypadki, mając pełną świadomość, że większość tego, co tu powiemy, okaże się jednak nieprawdą.

W pierwszym rozdziale wspomnieliśmy o liście 23 wielkich problemów, którą David Hilbert przedstawił w 1900 roku. Większość z nich udało się rozwiązać i wydaje się, że jego apel: „Musimy wiedzieć. Dowiemy się" spotkał się z szerokim odzewem. Hilbert powiedział również, że „w matematyce nie ma mowy o *ignorabimus*"[91], a jednak Kurt Gödel obalił tę wiarę w nasze możliwości poznawcze, przedstawiając dowód twierdzenia o niezupełności, z którego wynika, że niektóre problemy matematyczne mogą nie mieć rozwiązania w ramach tradycyjnego logicznego formalizmu matematyki. Nie chodzi o to, że rozwiązanie takich kwestii jest niemożliwe, jak dokonanie kwadratury koła, ale raczej o to, że mogą one być nierozstrzygalne – że nie da się ich ani udowodnić, ani obalić. Niewykluczone, że taki los czeka niektóre z obecnych nierozwiązanych wielkich problemów. Zdziwiłbym się, gdyby hipoteza Riemanna należała do tej kategorii, a jeszcze bardziej bym się zdumiał, gdyby komuś udało się udowodnić, że hipoteza ta jest nierozstrzygalna, nawet gdyby faktycznie tak było. Równocześnie wydaje mi się, że problem P/NP może okazać się nierozstrzygalny lub trafić do jakiejś innej grupy zagadnień o mądrej

nazwie, która będzie wyszukanym synonimem stwierdzenia, że „tego się nie da zrobić”. Problem ten ma taką właśnie… no cóż, *aurę*.

Podejrzewam, że pod koniec XXI wieku będziemy już mieli dowody poprawności hipotez Riemanna, Bircha i Swinnertona-Dyera oraz luki masowej, a także udowodnimy błędność hipotezy Hodge’a i hipotezy o regularności rozwiązań równania Naviera–Stokesa w trzech wymiarach. Wydaje mi się też, że w 2100 roku problem P/NP wciąż będzie nierozwiązany, ale uda się go wyjaśnić w drugiej połowie XXII wieku. O ile jednak znam swoje szczęście, pewnie jutro dowiemy się, że ktoś dowiódł niepoprawności hipotezy Riemanna, a tydzień później ukaże się dowód, że klasa P różni się od NP.

Przejdźmy na pewniejszy grunt bardziej ogólnych obserwacji, które można wysunąć na podstawie wniosków płynących z historii. Jestem niemal pewny, że gdy uda się w końcu rozwiązać siedem problemów milenijnych, wiele z nich będzie już tylko niezbyt interesującymi historycznymi ciekawostkami. „Och, oni wtedy sądzili, że *to* jest ważne, prawda?”. Właśnie taki los spotkał niektóre problemy z listy Hilberta. Jestem również przekonany, że w ciągu najbliższych 50 lat powstanie kilka ważnych gałęzi matematyki, o których istnieniu obecnie nie mamy pojęcia. Okaże się wtedy, że nieliczne proste przykłady i niektóre podstawowe twierdzenia z tych działów istniały już od dawna, ale nikt nie zdawał sobie sprawy, iż te odrębne fragmenty świadczą o istnieniu głębokich i ważnych nowych dziedzin. Tak właśnie było w wypadku teorii grup, algebry macierzy, fraktali i chaosu. Nie mam wątpliwości, że tego typu sytuacja się powtórzy, ponieważ jest to jeden ze standardowych sposobów rozwoju matematyki.

Te nowe obszary ukształtują się za sprawą dwóch głównych czynników. Wyłonią się ze struktury wewnętrznej samej matematyki lub pojawią się w odpowiedzi na nowe pytania dotyczące otaczającego nas świata – nierzadko oba te czynniki działają jednocześnie. Podobnie jak w wypadku trzyetapowego procesu rozwiązywania problemów, który zaproponował Poincaré – przygotowanie, dojrzewanie i olśnienie – związek między matematyką i jej zastosowaniami nie polega na pojedynczym przejściu: nauka przedstawia problem, matematyka go rozwiązuje,

sprawa zamknięta. Przeciwnie, mamy do czynienia ze skomplikowaną siecią wymiany pytań i koncepcji: nowe rozwiązania matematyczne są inspiracją do przeprowadzenia dalszych doświadczeń i obserwacji lub snucia kolejnych teorii, a to z kolei napędza rozwój matematyki. Każdy z węzłów tej sieci, jeśli przyjrzeć mu się z bliska, okazuje się mniejszą siecią takiego samego rodzaju.

Obecnie matematyka ma styczność z większą liczbą dziedzin niż kiedykolwiek wcześniej. Do niedawna w matematyce głównym zewnętrznym źródłem inspiracji były nauki fizyczne. Również kilka innych dziedzin odgrywało pewną rolę: biologia i socjologia przyczyniły się do rozwoju rachunku prawdopodobieństwa i statystyki, a filozofia wywarła duży wpływ na logikę matematyczną. W przyszłości możemy się spodziewać coraz większego wkładu ze strony biologii, medycyny, informatyki, ekonomii, socjologii, a bardzo możliwe, że również polityka, przemysł filmowy i sport odegrają tu pewną rolę. Podejrzewam, że kolejne wielkie problemy pojawią się na gruncie biologii, ponieważ związek między matematyką i biologią jest już dość mocny. Jednym z istotnych czynników w tym wypadku jest gwałtowny rozwój naszych możliwości gromadzenia danych na temat biologii i biochemii. Potrafimy już na przykład przeprowadzić sekwencjonowanie niewielkiego genomu za pomocą urządzenia wielkości pendrive'a, którego działanie opiera się na wykorzystaniu nanoporów. Dzięki zastosowaniu tych lub innych rozwiązań technicznych, których większość już istnieje, już niedługo równie łatwo będzie można dokonać analizy dużych genomów.

Procesy te mogą doprowadzić do poważnych zmian, ale musimy najpierw znaleźć lepsze metody pozyskiwania wiedzy ze zgromadzonych danych. W biologii dane same w sobie nie mają większego znaczenia. Liczy się proces. Procesem jest ewolucja, a także podział komórki, rozwój embrionu, zachorowanie na raka, ruch tłumu, działanie mózgu i dynamika globalnego ekosystemu. Najlepszy znany nam obecnie sposób na ustalenie przebiegu procesu na podstawie jego najważniejszych składników polega na wykorzystaniu matematyki. Pojawiają się zatem ważne pytania nowego rodzaju: w jaki sposób przebiega proces dynamiczny

w obecności skomplikowanej, ale dobrze określonej informacji porządkującej (ciągów DNA)? W jaki sposób działanie środowiska i zmiany genetyczne ograniczają ewolucję? Jak reguły decydujące o wzroście komórki, jej podziale, ruchliwości, lepkości i śmierci sprawiają, że rozwijający się organizm zyskuje określony kształt? W jaki sposób przepływ elektronów i związków chemicznych w sieci komórek nerwowych wpływa na to, co organizm może postrzegać i jak jest w stanie zareagować?

Kolejnym, sprawdzonym już źródłem inspiracji dla matematyki jest informatyka. Zwykle uważa się ją za narzędzie umożliwiające uprawianie matematyki, ale tak samo matematykę można postrzegać jako narzędzie umożliwiające zrozumienie i ustrukturyzowanie procesu obliczeniowego. Ta obustronna wymiana nabiera coraz większego znaczenia dla dobrej kondycji i rozwoju obu dziedzin i niewykluczone, że kiedyś może dojść do ich połączenia. Niektórzy matematycy uważają, że w ogóle nie należało dopuścić do ich rozdzielenia. Spośród wielu zagadnień z tego zakresu najszybciej przychodzi nam na myśl ponownie problem wielkich zbiorów danych. Nie chodzi tu tylko o wspomniane przed chwilą sekwencjonowanie DNA, ale także dane związane z przewidywaniem trzęsień ziemi, ewolucją, globalnym klimatem, obrotem papierami wartościowymi, międzynarodowym rynkiem finansowym i nowymi technologiami. Problem polega na wykorzystaniu olbrzymich ilości danych do sprawdzenia i udoskonalenia modeli matematycznych świata rzeczywistego w taki sposób, byśmy mogli sprawować faktyczną kontrolę nad bardzo skomplikowanymi zjawiskami.

Najpewniejsza moim zdaniem prognoza jest pod pewnym względem negacją, ale można również widzieć w niej potwierdzenie ciągłej kreatywności społeczności matematyków. Wszyscy aktywnie działający matematycy mają od czasu do czasu wrażenie, że badane przez nich zagadnienie żyje własnym życiem. Problemy rozwiązują się tak, jak chce tego matematyka, a nie matematycy. Wybieramy pytania, które chcemy postawić, ale nie mamy wpływu na to, jakie uzyskamy odpowiedzi. To poczucie ma związek z dwiema głównymi szkołami myślenia na temat natury matematyki. Platonicy uważają, że „idealne byty” matematyczne wiodą niezależną egzystencję

„gdzieś tam”, w jakimś świecie innym od naszego świata fizycznego. (Można to samo wyrazić w bardziej subtelny sposób i zapewne będzie wówczas brzmiało bardziej sensownie, ale na tym polega sedno tego stanowiska). Dla innych matematyka jest wspólnym dziełem ludzkości, które w odróżnieniu od większości tego typu osiągnięć – jak system prawny i walutowy, etyka, moralność – jest tworem opartym na solidnym szkielecie logiki. Istnieją ściśle określone zasady mówiące, jakimi twierdzeniami wolno – a jakimi nie – dzielić się z innymi uczonymi. To właśnie te reguły sprawiają, że wydaje nam się, iż matematyka działa według własnego planu, a w umysłach matematyków rodzi się przeświadczenie, że *istnieje* ona gdzieś poza domeną ludzkiej aktywności. Moim zdaniem platonizm nie opisuje, czym jest matematyka. Jest to raczej opis tego, jak *odczuwają* matematykę ludzie zajmujący się tą dziedziną. Przypomina to intensywne odczucie „czerwieni”, którego doświadczamy, patrząc na różę, krew lub światło drogowe. Filozofowie określają takie odczucia łacińskim słowem *qualia* (l. poj. *quale*), a niektórzy uważają, że nasze przekonanie o istnieniu wolnej woli jest w istocie przejawem *qualis* wynikającego z tego, jak mózg podejmuje decyzje. Gdy zastanawiamy się nad różnymi możliwościami, czujemy, że mamy prawdziwą swobodę dokonania wyboru – bez względu na to, czy dynamika mózgu jest faktycznie w pewnym sensie deterministyczna. Podobnie platonizm jest *quali* brania udziału we wspólnym przedsięwzięciu ludzkości, rozwijanym zgodnie ze ściśle określonymi zasadami logicznej dedukcji.

Zatem może nam się wydawać, że matematyka żyje własnym życiem, nawet jeśli powstaje w wyniku wysiłku umysłowego całej ludzkości. Historia uczy nas, że działanie tak rozumianej matematyki jest bardziej innowacyjne i zaskakujące od wszystkiego, co mógłby przewidzieć pojedynczy człowiek. Wszystko to prowadzi okrężną drogą do głównej tezy, jaką chciałem tu postawić: jedynym, co możemy z całą pewnością powiedzieć na temat przyszłości matematyki, jest to, że będzie ona nieprzewidywalna. Najważniejsze pytania matematyczne następnego stulecia pojawią się jako naturalne, a może nawet nieuniknione konsekwencje powiększenia się naszej wiedzy na temat zagadnień, które obecnie uważamy za

największe problemy matematyki. Niemal z całą pewnością będą to jednak pytania, których obecnie nie możemy sobie nawet wyobrazić. I bardzo dobrze – powinniśmy się z tego cieszyć.

89 Mówiąc to, Bohr mógł mieć na myśli pewną istotną kwestię. Teorie naukowe sprawdza się, weryfikując ich przewidywania, ale tylko w nielicznych wypadkach dotyczą one przyszłości. W większości sytuacji mamy do czynienia ze stwierdzeniami warunkowymi typu: „jeśli przepuścimy światło przez pryzmat, rozdzieli się ono na różne barwy". Takie „przewidywanie" nie mówi, kiedy to nastąpi. Zatem, paradoksalnie, możemy wysuwać przewidywania na temat pogody, wcale jej nie prognozując. „Jeśli ciepłe powietrze z cyklonu zetknie się z prądem zimnego powietrza, to będzie padał śnieg" – to jest naukowe przewidywanie, ale nie prognoza pogody.

90 Cytat ten w różnym brzmieniu przypisuje się około trzydziestu różnym osobom, między innymi Samowi Goldwynowi, Woody'emu Allenowi, Winstonowi Churchillowi i Konfucjuszowi. Zob.: http://www.larry.denenberg.com/predictions.html.

91 Jest to nawiązanie do łacińskiej maksymy *Ignoramus et ignorabimus* – nie wiemy i nie dowiemy się – sformułowanej przez niemieckiego zoologa Emila du Bois-Reymonda (przyp. tłum.).

17. Dwanaście na przyszłość

Nie chciałbym, żeby czytelnicy tej książki odnieśli wrażenie, że większość problemów matematycznych została już rozwiązana i do rozstrzygnięcia pozostało jeszcze tylko kilka ostatnich, naprawdę trudnych kwestii. Badania matematyczne przypominają odkrywanie nowego kontynentu. W miarę jak poszerza się obszar naszej wiedzy, granica z niezbadanym jeszcze terenem ciągle się wydłuża. Nie chcę przez to powiedzieć, że im więcej odkrywamy matematyki, tym mniej wiemy, ale raczej to, że im głębiej ją poznajemy, tym lepiej uświadamiamy sobie, czego jeszcze nie wiemy. Jednak to, czego nie wiemy, zmienia się z upływem czasu – istniejące od dawna problemy znikają, a ich miejsce zajmują nowe zagadki. Natomiast nasza wiedza stale się powiększa – z wyjątkiem sytuacji, gdy od czasu do czasu ktoś gubi swoje notatki.

Abyśmy mogli uzmysłowić sobie, czego obecnie *nie wiemy*, poza omówionymi tu wielkimi problemami, pozwolę sobie w tym rozdziale przedstawić dwanaście nierozstrzygniętych kwestii, które od jakiegoś już czasu nie dają matematykom spokoju. Wybierając te problemy, kierowałem się tym, żeby były to pytania łatwe do zrozumienia, co, jak już wielokrotnie się przekonaliśmy, nie oznacza wcale, że równie łatwo można znaleźć na nie odpowiedzi. Być może kiedyś niektóre z tych kwestii trafią na jakąś listę wielkich problemów – będzie to zależało głównie od metod, jakie trzeba wymyślić, żeby je rozwiązać, i od tego, do czego pytania te prowadzą, a nie od samych rozwiązań.

Problem Brocarda

Dla dowolnej liczby naturalnej n jej silnia, $n!$, jest iloczynem

$$n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Wartość ta jest równa liczbie różnych możliwych sposobów ustawienia po kolei *n* obiektów. Na przykład 32 litery polskiego alfabetu można ułożyć na

$$32! = 263\,130\,836\,933\,693\,530\,167\,218\,012\,160\,000\,000$$

różnych sposobów. W artykułach napisanych w latach 1876 i 1885 Henri Brocard zauważył, że wartości

$$4! + 1 = 24 + 1 = 25 = 5^2,$$
$$5! + 1 = 120 + 1 = 121 = 11^2,$$
$$7! + 1 = 5040 + 1 = 5041 = 71^2$$

są kwadratami. Nie udało mu się znaleźć innych liczb, których silnia powiększona o 1 byłaby kwadratem, i postawił pytanie, czy liczby takie, poza tymi trzema przypadkami, w ogóle istnieją. Niezależnie od niego to samo pytanie zadał indyjski geniusz matematyczny i samouk Srinivasa Ramanujan w 1913 roku. W 2000 roku Bruce Berndt i William Galway udowodnili z wykorzystaniem komputera, że innych liczb tego typu nie ma w przedziale od 1 do miliarda.

Nieparzyste liczby doskonałe

Liczba doskonała jest równa sumie wszystkich swoich dzielników właściwych (to znaczy mniejszych od niej liczb, przez które dzieli się bez reszty). Przykładami takich liczb są

$$6 = 1 + 2 + 3,$$
$$28 = 1 + 2 + 4 + 7 + 14.$$

Euklides udowodnił, że jeśli $2^n - 1$ jest liczbą pierwszą, to $2^{n-1}\left(2^n - 1\right)$ jest liczbą doskonałą. Przytoczone przed chwilą przykłady odpowiadają *n* = 2, 3. Tego typu liczby pierwsze nazywamy liczbami pierwszymi Mersenne'a. Na razie znamy 47 tego typu liczb, z których największa, $2^{43\ 112\ 609} - 1$, jest również największą znaną liczbą pierwszą[92]. Euler udowodnił, że wszystkie liczby doskonałe muszą mieć taką postać, ale jak dotąd nikomu nie udało się znaleźć nieparzystej liczby doskonałej ani dowieść, że nieparzyste liczby tego typu nie mogą istnieć. Carl Pomerance przedstawił nieścisłą argumentację, z której wynika, że liczby takie nie istnieją. Dowolna nieparzysta liczba doskonała, jeśli istnieje, musi spełniać bardzo rygorystyczne warunki. Nie może być mniejsza niż 10^{300}, musi mieć czynnik pierwszy większy od 10^8, jej drugi pod względem wielkości czynnik pierwszy musi wynosić co najmniej 10^4 i musi mieć przynajmniej 75 czynników pierwszych, w tym przynajmniej 12 z nich musi być różnych.

Hipoteza Collatza

Weźmy dowolną liczbę naturalną. Jeśli jest parzysta, podzielmy ją przez 2. Jeśli jest nieparzysta, pomnóżmy ją przez 3 i dodajmy 1. Teraz tak samo postąpmy z wynikiem tej operacji i powtarzajmy całą procedurę bez końca. Co się stanie?

Weźmy na przykład liczbę 12. Kolejne liczby to

$$12 \to 6 \to 3 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1,$$

po czym sekwencja $4 \to 2 \to 1 \to 4 \to 2 \to 1$ powtarza się bez końca. Hipoteza Collatza głosi, że ten sam ciąg końcowy pojawi bez względu na to, jaką liczbę wybierzemy na początku. Najczęściej w nazwie tej hipotezy pojawia się nazwisko Lothara Collatza, który wysunął ją w 1937 roku, ale występuje ona również pod wieloma

innymi nazwami: hipoteza $3n + 1$, hipoteza Ulama, problem Kakutaniego, hipoteza Thwaitesa, algorytm Hassego czy problem syrakuzański.

Cała trudność w badaniu tego zagadnienia polega na tym, że liczby czasami wybuchają. Jeśli na przykład wyjdziemy od liczby 27, to taki ciąg wzrośnie aż do wartości 9232, ale potem i tak w końcu zmniejszy się do 1 po 111 krokach. Dzięki symulacjom komputerowym udało się potwierdzić tę hipotezę dla wszystkich liczb od 1 do $5{,}764 \times 10^{18}$. Udowodniono również, że nie istnieje żaden inny cykl oprócz $4 \rightarrow 2 \rightarrow 1$, który obejmowałby mniej niż 35 400 liczb. Nie udało się wykluczyć możliwości, że jakaś liczba może prowadzić do ciągu zawierającego coraz większe liczby rozdzielone mniejszymi wartościami. Ilia Krasikov i Jeffrey Lagarias udowodnili, że w przedziale od 1 do n określony odsetek liczb, wynoszący co najmniej: stała $\times n^{0,84}$, osiągnie ostatecznie liczbę 1. Zatem wyjątki, jeśli istnieją, muszą być rzadkie[93].

Istnienie prostopadłościanów doskonałych

Ten problem polega na przeniesieniu zagadnienia istnienia trójek pitagorejskich i wzoru na nie do trzech wymiarów. Cegiełka Eulera jest prostopadłościanem – czyli bryłą w kształcie cegły – którego krawędzie i przekątne wszystkich ścian mają długości wyrażające się liczbami naturalnymi. Najmniejszą cegiełkę Eulera odkrył Paul Halcke w 1719 roku. Jej krawędzie mają długości 240, 117 i 4, natomiast przekątne ścian wynoszą: 267, 244 i 125. Euler przedstawił wzory na takie prostopadłościany, podobne do wzoru na trójki pitagorejskie, ale nie wynikają z nich wszystkie możliwe rozwiązania.

Nie wiadomo, czy prostopadłościan doskonały w ogóle istnieje, czyli innymi słowy, czy istnieje cegiełka Eulera, której główna przekątna przechodząca przez jej wnętrze (łącząca dwa przeciwległe rogi) również miałaby długość wyrażającą się liczbą naturalną. (Istnieją cztery takie przekątne, ale wszystkie mają tę samą

długość). Wiadomo, że wzory Eulera nie pozwalają otrzymać przykładu takiego prostopadłościanu. Taka cegiełka, jeśli istnieje, musi spełniać kilka warunków – na przykład długości jej krawędzi muszą być wielokrotnościami liczb 5, 7, 11 oraz 19 i dla każdej z tych liczb musi istnieć przynajmniej jedna krawędź, której długość dzieli się przez nią bez reszty. Z badań komputerowych wynika, że jedna z krawędzi musi mieć długość wynoszącą przynajmniej bilion.

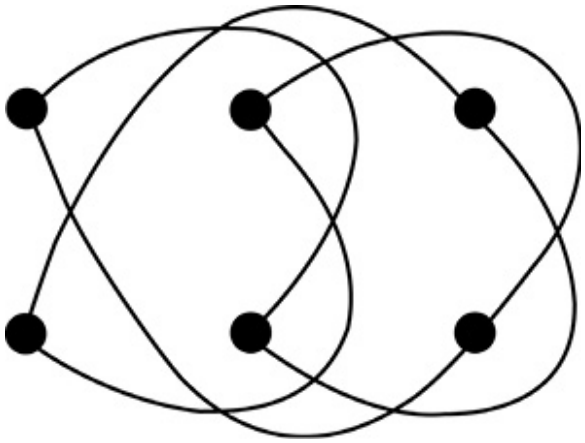
Znamy kilka przykładów bliskich doskonałości. Cegiełka o krawędziach 672, 153 i 104 ma główną przekątną, której długość wyraża się liczbą naturalną, i dwie z trzech przekątnych jej ścian również są liczbami naturalnymi. W 2004 roku Jorge Sawyer i Clifford Reiter udowodnili, że istnieją doskonałe równoległościany[94]. Równoległościan jest bryłą przypominającą prostopadłościan, z tą różnicą, że jej ściany są równoległobokami, a więc jest wykrzywiony. Krawędzie doskonałego równoległościanu mają długości: 271, 106 i 103; krótsze przekątne jego ścian mają długości: 101, 266 i 255; dłuższe przekątne jego ścian mają długości: 183, 312 i 323; natomiast przekątne całej bryły mają długości 374, 300, 278 i 272.

Hipoteza samotnego biegacza

Ten problem wywodzi się ze skomplikowanego działu matematyki znanego jako teoria aproksymacji diofantycznej. Sformułował go Jörg Wills w 1967 roku. Tak obrazową nazwę nadał temu problemowi Luis Goddyn w 1998 roku. Przyjmijmy, że n zawodników biegnie ze stałą szybkością po torze kołowym o jednostkowej długości, z tym że u każdego z biegaczy tempo to jest inne. Czy każdy z biegaczy będzie kiedyś samotny – to znaczy, czy nastąpi taka chwila, w której od pozostałych zawodników będzie go dzieliła odległość większa niż $1/n$? Oczywiście poszczególni biegacze będą samotni w różnych momentach. Hipoteza głosi, że odpowiedź na tak postawione pytanie zawsze brzmi: „tak”, i udało się to udowodnić dla $n = 4, 5, 6$ oraz 7.

Hipoteza Conwaya

Hipoteza Conwaya dotyczy szczególnego rodzaju sieci (grafu) na płaszczyźnie, która charakteryzuje się tym, że każde dwie krawędzie spotykają się ze sobą dokładnie raz (zob. ryc. 48). Mogą się spotykać albo we wspólnym wierzchołku (węźle), albo się przecinać, ale nie jedno i drugie. Jeśli się przecinają, to każda z krawędzi musi przejść na drugą stronę przecinanej linii (to znaczy nie mogą być styczne). W nieopublikowanej pracy John Horton Conway wysunął hipotezę, że w każdej tego typu sieci liczba krawędzi musi być mniejsza lub równa liczbie wierzchołków. W 2011 roku Radoslav Fulek i János Pach udowodnili, że każdy taki graf z *n* wierzchołkami ma najwyżej 1,428*n* krawędzi[95].



Ryc. 48. Sieć spełniająca wymagania hipotezy Conwaya

Niewymierność stałej Eulera

Nie znamy żadnego „jawnego” wzoru na wartość szeregu harmonicznego

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots \frac{1}{n}$$

i zapewne wzór taki w ogóle nie istnieje. Istnieje jednak pewne doskonałe przybliżenie: gdy *n* rośnie, H_n zbliża się do wartości $\ln n + \gamma$. Symbol γ oznacza tu stałą Eulera, której wartość liczbowa wynosi w przybliżeniu 0,5772156649. Euler wyprowadził ten wzór w 1734 roku, a Lorenzo Mascheroni badał występującą w nim stałą w 1790 roku. Ani jeden, ani drugi uczony nie posługiwał się symbolem γ.

Stała Eulera jest jedną z tych dziwnych liczb spotykanych w matematyce, takich jak π i *e*, które pojawiają się w różnych miejscach, a mimo to zdają się wieść niezależny żywot i nie dają się wyrazić w żaden elegancki sposób za pomocą jakichś prostszych wartości. W rozdziale 3 mówiliśmy o tym, że π i *e* są liczbami przestępnymi, czyli że nie są rozwiązaniami żadnego równania algebraicznego o całkowitych współczynnikach. W szczególności są to liczby niewymierne – nie można ich przedstawić za pomocą ilorazu dwóch liczb całkowitych. Powszechnie uważa się, że stała Eulera również jest liczbą przestępną, ale tak naprawdę nie wiemy nawet, czy jest niewymierna. Jeśli $\gamma = p/q$, dla całkowitych liczb *p* i *q*, to *q* musi wynosić przynajmniej $10^{242\,080}$.

Stała Eulera odgrywa ważną rolę w wielu działach matematyki, od funkcji dzeta Riemanna po kwantową teorię pola. Pojawia się w wielu kontekstach i występuje w różnych wzorach. To okropne, że nie potrafimy powiedzieć nawet, czy jest wymierna.

Rzeczywiste ciała kwadratowe

W rozdziale 7 przekonaliśmy się, że niektóre algebraiczne ciała liczbowe cechują się jednoznacznością rozkładu na czynniki pierwsze, a niektóre nie. Najlepiej poznanymi ciałami liczbowymi są

ciała kwadratowe, które otrzymujemy, wyciągając pierwiastek kwadratowy z wybranej liczby *d*; nie jest ona kwadratem liczby całkowitej, a nawet nie ma czynników pierwszych będących kwadratami. Odpowiedni pierścień liczb algebraicznych całkowitych składa się ze wszystkich liczb postaci $a + b\sqrt{d}$, gdzie *a* i *b* są liczbami całkowitymi, gdy *d* nie jest liczbą postaci 4*k* + 1, jeżeli natomiast *d* ma taką postać, to *a* i *b* mogą być albo liczbami całkowitymi, albo nieparzystymi liczbami całkowitymi podzielonymi przez 2 (wtedy i *a*, i *b* muszą mieć taką postać).

Wiadomo, że gdy *d* jest liczbą ujemną, to rozkład na czynniki pierwsze jest jednoznaczny dla dokładnie dziewięciu wartości: –1, –2, –3, –7, –11, –19, –43, –67 i –163. Udowodnienie jednoznaczności rozkładu jest wówczas dość oczywiste, ale ustalenie, czy mogą istnieć jeszcze inne tego typu przypadki, okazuje się znacznie trudniejsze. W 1934 roku Hans Heilbronn i Edward Linfoot udowodnili, że lista ta może się powiększyć najwyżej o jeszcze jedną liczbę. Kurt Heegner w 1952 roku przedstawił dowód, że lista ta jest pełna, ale inni matematycy sądzili, że w jego dowodzie jest luka. W 1967 roku Harold Stark przedstawił pełny dowód, zauważając jednocześnie, że nie różni się on zasadniczo od dowodu Heegnera – to znaczy okazało się, że owa luka była bez znaczenia. Mniej więcej w tym samym czasie Alan Baker znalazł inny dowód kompletności tej listy.

Z zupełnie inną sytuacją mamy natomiast do czynienia, gdy *d* jest liczbą dodatnią. Wtedy rozkład jest jednoznaczny dla dużo większej liczby wartości *d*. W przedziale od 1 do 50 rozkład jest jednoznaczny dla: 2, 3, 5, 6, 7, 11, 13, 14, 17, 19, 21, 22, 23, 29, 31, 33, 37, 38, 41, 43, 46 i 47. Obliczenia komputerowe pozwalają znaleźć znacznie więcej takich wartości. O ile nam wiadomo, może istnieć nieskończenie wiele dodatnich wartości *d*, dla których odpowiednie ciało kwadratowe cechuje się jednoznacznością rozkładu na czynniki. Cohen i Lenstra przeprowadzili analizę heurystyczną, z której wynika, że mniej więcej trzy czwarte wszystkich dodatnich wartości *d* powinno prowadzić do powstania pierścieni o jednoznacznym

rozkładzie. Wyniki obliczeń komputerowych potwierdzają te szacunki. Problem polega na udowodnieniu, że obserwacje te są poprawne.

Mrówka Langtona

W XXI wieku coraz wyraźniej widać, że niektóre tradycyjne metody modelowania matematycznego nie są w stanie sprostać złożoności problemów stojących przed ludzkością, takich jak globalny system finansowy, dynamika ekosystemów i rola genów we wzroście organizmów żywych. Tego typu układy zawierają często olbrzymią liczbę oddziałujących ze sobą elementów – ludzi, przedsiębiorstw, organizmów czy genów. Nierzadko takie oddziaływania można dość dokładnie modelować za pomocą prostych reguł. W związku z tym w ciągu ostatnich 30 lat wykształcił się nowy rodzaj modeli umożliwiających przeprowadzenie bezpośredniej analizy zachowania układów z wieloma elementami. Aby na przykład zrozumieć, jak na stadionie sportowym przemieszcza się tłum składający się ze 100 000 osób, badacze nie uśredniają wszystkich ludzi w celu stworzenia swego rodzaju ludzkiej cieczy i zbadania, jak ona przepływa. Zamiast tego konstruują model komputerowy obejmujący 100 000 elementów, definiują odpowiednie reguły i uruchamiają symulację, która pokaże, jak zachowuje się taki komputerowy tłum. Tego rodzaju modele nazywamy układami złożonymi.

Przekonajmy się, jak wygląda ten fascynujący nowy obszar matematyki, omawiając jeden z najprostszych układów złożonych i wyjaśniając, dlaczego wciąż w pełni go nie rozumiemy. Chodzi o tak zwaną mrówkę Langtona. Christopher Langton jest jednym z pierwszych członków Instytutu Santa Fe, założonego w 1984 roku przez George'a Cowana, Murraya Gell-Manna i innych uczonych w celu promowania teorii i zastosowań praktycznych układów złożonych. Langton wymyślił swoją mrówkę w 1986 roku. Jest ona szczególnym rodzajem automatu komórkowego, czyli układu komórek, które tworzą kwadratową sieć i mogą przyjmować różne stany oznaczane różnymi kolorami. W każdym kroku czasowym kolor

każdej komórki zmienia się w zależności od tego, jakie kolory mają jej sąsiadki.

Obowiązujące tu reguły są absurdalnie proste. Mrówka żyje w nieskończenie dużej kwadratowej sieci komórek, które początkowo wszystkie są białe. Niesie ze sobą dwie niewyczerpywalne puszki szybko schnącej farby w kolorze czarnym i białym. Może być zwrócona na północ, południe, wschód lub zachód. Przyjmijmy, że na początku jest zwrócona na północ – wybranie innego kierunku spowoduje powstanie takiego samego, tylko odpowiednio obróconego wzoru. W każdym kroku nasza mrówka sprawdza, jaki kolor ma komórka, w której się znajduje, i zmienia go z białego na czarny lub z czarnego na biały. Jeśli komórka była biała, obraca się następnie o 90 stopni w prawo i przechodzi o jedną komórkę do przodu. Jeśli natomiast była czarna, obraca się 90 stopni w lewo i robi krok w przód. Następnie powtarza te czynności bez końca.

Jeśli przeprowadzimy symulację zachowania takiej mrówki[96], to na początku będzie ona tworzyła proste, dość symetryczne wzory czarnych i białych kwadratów. Od czasu do czasu będzie powracała do kwadratu, który już odwiedziła, ale nie spowoduje to powtórzenia tych samych czynności, ponieważ teraz kolor tej komórki będzie inny niż przy pierwszej wizycie, za drugim razem więc uda się w przeciwnym kierunku. W miarę rozwoju symulacji wzór tworzony przez mrówkę staje się chaotyczny i przypadkowy. Na tym etapie moglibyśmy zupełnie zasadnie przyjąć, że takie chaotyczne zachowanie będzie się utrzymywało do końca symulacji, bez względu na to, jak długo by trwała. W końcu, gdy mrówka wróci ponownie do chaotycznego obszaru, wykona zapewne chaotyczny ciąg obrotów i przemalowań. Jeśli poprowadzimy tę symulację przez kolejne mniej więcej 10 000 kroków, zachowanie mrówki będzie się zdawało potwierdzać ten wniosek. Jeżeli jednak symulacji nie przerwiemy w tym miejscu, pojawi się pewien wzór. Mrówka zacznie powtarzać cykl 104 kroków. Po zakończeniu każdego takiego cyklu przesuwa się po przekątnej o dwie komórki. W efekcie maluje ukośny pas czarnych i białych komórek, zwany autostradą, który ciągnie się bez końca (zob. ryc. 49).

Ryc. 49. Autostrada namalowana przez mrówkę Langtona

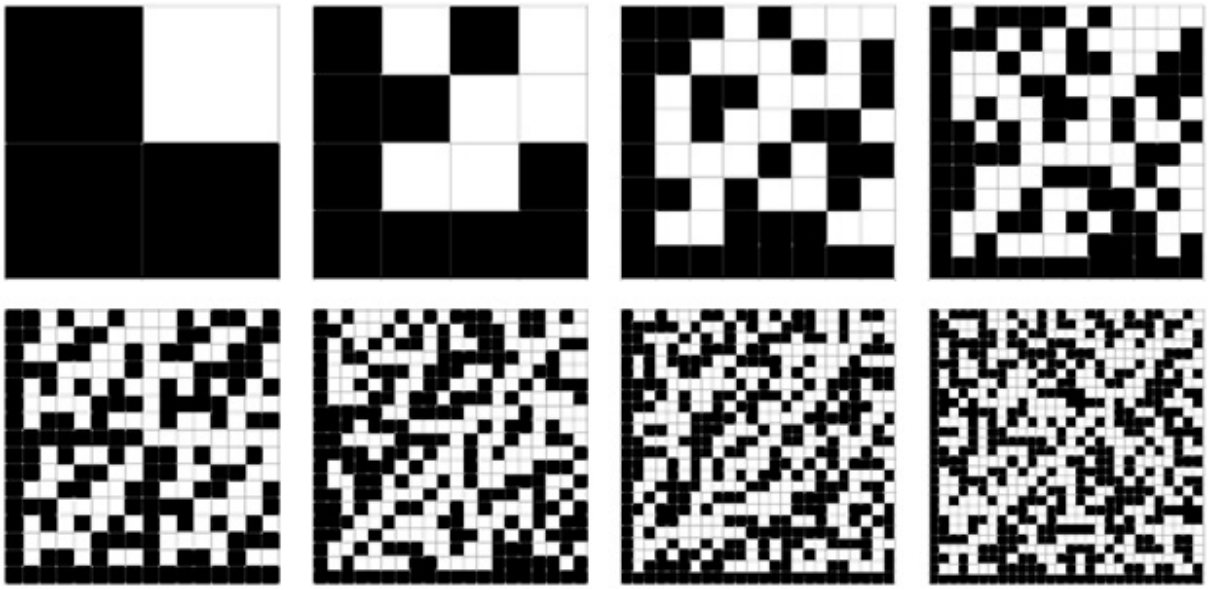
Wszystko, co dotąd powiedzieliśmy, można udowodnić z zachowaniem pełnej ścisłości, podając po prostu listę kroków, które wykonuje mrówka. Taki dowód byłby dość długi – dotyczyłby w końcu listy 10 000 kroków – ale całkowicie poprawny. Sytuacja robi się bardziej interesująca, gdy postawimy nieco ogólniejsze pytanie. Załóżmy, że przed uruchomieniem mrówki pomalowaliśmy skończoną liczbę komórek na czarno. Kwadraty do pomalowania możemy wybrać w zupełnie dowolny sposób: mogą one tworzyć przypadkowe kropki, czarny prostokąt, a nawet podobiznę Mony Lizy. Możemy zamalować milion lub miliard kwadratów, ale nie nieskończenie wiele. Co się wtedy stanie?

Na początku, gdy mrówka natrafia na pomalowane przez nas czarne kwadraty, jej poczynania będą zupełnie inne niż poprzednio. Będzie się wałęsała po całym obszarze, rysując zawiłe wzory, a potem przemalowywała je na nowo… Jednak w każdej symulacji, jaką dotychczas przeprowadzono, mrówka przechodzi w końcu do budowania swojej autostrady, powtarzając zawsze ten sam cykl o 104 krokach. Czy tak będzie zawsze? Czy autostrada jest jedynym „atraktorem” dynamiki mrówki? Nie wiadomo. Jest to jeden z podstawowych nierozstrzygniętych problemów teorii złożoności. Udało się jedynie ustalić, że bez względu na to, jak wygląda

początkowa konfiguracja czarnych komórek, mrówka nie może pozostać na zawsze uwięziona w jakimś ograniczonym obszarze siatki.

Hipoteza o istnieniu macierzy Hadamarda

Macierz Hadamarda, nazwana tak na cześć francuskiego matematyka Jacques'a Hadamarda, jest kwadratową tablicą wypełnioną zerami i jedynkami w taki sposób, że dowolne dwa wiersze i dowolne dwie kolumny mają połowę wartości jednakowych, a połowę różnych. Jeśli kratki z jedynką oznaczymy kolorem czarnym, a komórki zawierające zera – białym, to macierz Hadamarda można przedstawić w sposób graficzny tak, jak to pokazano na rycinie 50 dla macierzy o rozmiarach 2, 4, 8, 12, 16, 20, 24 i 28. Macierze te pojawiają się wielu problemach matematycznych i informatycznych, szczególnie w teorii kodowania. (W niektórych zastosowaniach, a także w oryginalnym ujęciu Hadamarda, białym kwadratom odpowiadają wartości –1, nie 0).



Ryc. 50. Macierze Hadamarda o rozmiarze 2, 4, 8, 12, 16, 20, 24 i 28 (http://mathworld.wolfram.com/HadamardMatrix.html)

Hadamard udowodnił, że takie macierze mogą istnieć tylko dla *n* = 2 oraz *n* będącego wielokrotnością liczby 4. Twierdzenie Paleya z 1993 roku dowodzi, że macierz Hadamarda na pewno istnieje, gdy jej rozmiar jest liczbą będącą wielokrotnością 4 postaci $2^a(p^b+1)$, gdzie *p* jest nieparzystą liczbą pierwszą. Twierdzenie to nie obejmuje następujących wielokrotności liczby 4: 92, 116, 156, 172, 184, 188, 232, 236, 260, 268 i innych większych wartości. Wysunięto jednak hipotezę, że macierz Hadamarda istnieje dla wszystkich wielokrotności 4. W 1985 roku Kazue Sawade znalazła macierz Hadamarda o rozmiarze 268. Dla pozostałych wymienionych przed chwilą wartości nieobjętych twierdzeniem Paleya również udało się odkryć odpowiednie macierze. W 2004 roku Hadi Kharaghani i Behruz Tayfeh-Rezaie przedstawili macierz Hadamarda o rozmiarze 428. Obecnie najmniejszym rozmiarem, dla którego nie wiadomo, czy istnieje macierz Hadamarda, jest wartość 668.

Równanie Fermata–Catalana

W tym problemie chodzi o równanie diofantyczne $x^a + y^b = z^c$, gdzie wykładniki *a*, *b* i *c* są liczbami naturalnymi. Będę je nazywał równaniem Fermata–Catalana, ponieważ jego rozwiązania mają związek zarówno z wielkim twierdzeniem Fermata (omówionym w rozdziale 7), jak i z hipotezą Catalana (o której mówiliśmy w rozdziale 6). Gdy wartości *a*, *b* i *c* są małe, istnienie niezerowych rozwiązań całkowitych nie jest specjalnie zaskakujące. Na przykład gdy wszystkie wykładniki są równe 2, mamy do czynienia z równaniem Pitagorasa, o którym już od czasów Euklidesa wiadomo, że ma nieskończenie wiele rozwiązań. Zatem najbardziej interesujące są przypadki, gdy wykładniki są duże. Ścisła definicja określenia „duże” brzmi: wartość $s = 1/a + 1/b + 1/c$ musi być

mniejsza od 1. Znamy tylko dziesięć rozwiązań równania Fermata–Catalana z dużymi wykładnikami:

$1 + 233^3 = 3^2$ $17^7 + 76\,271^3 = 21\,063\,928^2$
$2^5 + 7^2 = 3^4$ $1414^3 + 2\,213\,459^2 = 65^7$
$7^3 + 13^2 = 2^9$ $9262^3 + 15\,312\,283^2 = 113^7$
$2^7 + 17^3 = 71^2$ $43^8 + 9622^3 = 30\,042\,907^2$
$3^5 + 11^4 = 122^2$ $33^8 + 159\,034^2 = 15\,613^3$

Pierwsze z nich uważa się za równanie z dużymi wykładnikami, ponieważ $1 = 1^a$ dla dowolnego a i wystarczy przyjąć $a = 7$, żeby spełnić definicję. Hipoteza Fermata–Catalana głosi, że jeśli wykładniki a, b i c spełniają definicję określenia „duże", to równanie Fermata–Catalana ma jedynie skończoną liczbę rozwiązań całkowitych nieposiadających wspólnego czynnika. Najważniejszą pracą związaną z tą hipotezą jest dowód, który Henri Darmon i Loïc Merel przedstawili w 1997 roku. Uczeni ci udowodnili, że równanie Fermata–Catalana nie ma rozwiązań, gdy $c = 3$, a wykładniki a i b są większe lub równe 3. To prawie wszystko, co wiemy na ten temat. Wydaje się, że dalsze postępy zależą od prawdziwości fascynującej nowej hipotezy, którą teraz omówimy.

Hipoteza ABC

W 1983 roku Richard Mason zauważył, że matematycy przeoczyli jeden przypadek wielkiego twierdzenia Fermata, a mianowicie: pierwszą potęgę. Chodzi zatem o równanie $a + b = c$.

Na pierwszy rzut oka taka uwaga wydaje się zupełnie bezsensowna. Nie trzeba być orłem z algebry, żeby rozwiązać to równanie, wyrażając dowolną z trzech zmiennych za pomocą pozostałych dwóch. Na przykład $a = c - b$. Wszystko się jednak zmienia, gdy rozważymy to równanie w odpowiednim kontekście.

Mason zauważył, że zagadnienie to nabiera głębi, gdy postawimy odpowiednie pytania na temat *a*, *b* i *c*. Wynikiem tej niezwykłej idei jest nowa hipoteza z zakresu teorii liczb, której konsekwencje sięgają bardzo daleko. Jej udowodnienie pozwoliłoby nam się uporać z wieloma nierozwiązanymi problemami oraz uzyskać lepsze i prostsze dowody ważnych twierdzeń z teorii liczb. Jest to tak zwana hipoteza ABC, opierająca się na luźnym związku między liczbami całkowitymi i wielomianami. Za jej poprawnością przemawia mnóstwo dowodów.

Euklides i Diofantos znali przepis na trójki pitagorejskie, który obecnie zapisujemy w postaci wzoru (mówiliśmy o tym w rozdziale 6). Czy tę samą sztuczkę można powtórzyć dla innych równań? W 1851 roku Joseph Liouville udowodnił, że tego typu wzór nie istnieje dla równania Fermata, w którym wykładniki są większe lub równe 3. Mason zastosował podobne rozumowanie do prostszego równania

$$a(x)+b(x)=c(x)$$

wiążącego ze sobą trzy wielomiany. To oburzający pomysł, ponieważ wszystkie rozwiązania można znaleźć przy użyciu podstawowej algebry. Główny wniosek jest jednak elegancki i wcale nie oczywisty: jeśli każdy z wielomianów ma czynnik będący kwadratem, sześcianem lub jakąś wyższą potęgą, to takie równanie nie ma rozwiązań.

Twierdzenia o wielomianach mają często swoje odpowiedniki mówiące o liczbach całkowitych. W szczególności wielomiany nieprzywiedlne odpowiadają liczbom pierwszym. Twierdzenie Masona o wielomianach również ma swój naturalny odpowiednik w zbiorze liczb całkowitych, który wygląda następująco: jeśli $a + b = c$, gdzie *a*, *b* i *c* są liczbami naturalnymi bez wspólnego dzielnika, to liczba czynników pierwszych każdej z liczb *a*, *b* oraz *c* jest mniejsza od liczby *różnych* czynników pierwszych iloczynu *abc*. Niestety, można przedstawić proste przykłady pokazujące, że takie twierdzenie jest nieprawdziwe. W 1985 roku David Masser i Joseph

Oesterlé zmodyfikowali to sformułowanie i zaproponowali nową wersję hipotezy, która nie stoi w sprzeczności z żadnymi znanymi przykładami. Niewykluczone, że hipoteza ABC jest obecnie jednym z najważniejszych otwartych pytań w teorii liczb[97]. Gdyby komuś udało się potwierdzić jej poprawność, to wiele głębokich i trudnych twierdzeń, udowodnionych w ostatnich dziesięcioleciach dzięki olbrzymiej wiedzy i wysiłkowi uczonych, z miejsca zyskałoby nowe, proste dowody. Inną konsekwencją byłoby udowodnienie hipotezy Marshalla Halla: różnica między kwadratem i sześcianem dowolnych liczb musi być dość duża. Kolejnym potencjalnym zastosowaniem hipotezy ABC jest rozwiązanie problemu Brocarda, o którym mówiliśmy na samym początku tego rozdziału. W 1993 roku Marius Overholt udowodnił, że jeśli hipoteza ABC jest prawdziwa, to istnieje jedynie skończona liczba rozwiązań równania Brocarda.

Jedna z najbardziej interesujących konsekwencji hipotezy ABC ma związek z hipotezą Mordella. Faltings udowodnił ją z wykorzystaniem skomplikowanych metod, ale to osiągnięcie miałoby znacznie większe znaczenie, gdybyśmy mieli jeszcze jedną informację: ograniczenie na wielkość rozwiązań. W takim wypadku istniałby algorytm umożliwiający znalezienie ich wszystkich. W 1991 roku Noam Elkies pokazał, że takie ulepszone twierdzenie Faltingsa wynika z pewnej szczególnej wersji hipotezy ABC, w której różne pojawiające się w niej stałe mają ograniczenia. Laurent Moret-Bailly dowiódł, że prawdziwe jest również twierdzenie odwrotne, a mianowicie iż pełna hipoteza ABC wynika z nałożenia odpowiednio silnych ograniczeń na rozmiar rozwiązań *tylko jednego* równania diofantycznego: $y^2 = x^5 - x$. Chociaż hipoteza ABC nie jest tak dobrze znana jak wiele innych nierozstrzygniętych kwestii, to bez wątpienia jest ona jednym z wielkich problemów matematyki. Andrew Granville i Thomas Tucker uważają, że rozstrzygnięcie kwestii jej prawdziwości miałoby „olbrzymie znaczenie dla naszego rozumienia teorii liczb. Udowodnienie jej lub wykazanie fałszywości byłoby czymś wspaniałym”[98].

92 Najświeższe informacje na ten temat można znaleźć na stronie internetowej http://primes.utm.edu/. [W styczniu 2013 roku Curtis Cooper odkrył czterdziestą ósmą liczbę Mersenne'a: $2^{57\ 885\ 161} - 1$, która jest obecnie największą znaną liczbą pierwszą (przyp. tłum.)].

93 Ilia Krasikov, Jeffrey C. Lagarias, *Bounds for the* 3*x* + 1 *problem using difference inequalities* (*Ograniczenia na problem* 3*x* + 1 *wynikające z nierówności różnicowych*), „Acta Arithmetica" 2003, tom 109, s. 237–258.

94 Jorge F. Sawyer, Clifford A. Reiter, *Perfect parallelepipeds exist* (*Doskonałe równoległościany istnieją*), arXiv:0907.0220, 2009.

95 Radoslav Fulek, János Pach, *A computational approach to Conway's thrackle conjecture* (*Podejście numeryczne do hipotezy Conwaya*), „Computational Geometry" 2011, tom 44, s. 345–355.

96 Zob.: http://pl.wikipedia.org/wiki/Mrówka_Langtona.

97 Hipoteza ABC brzmi następująco: dla każdej wartości $\varepsilon > 0$ istnieje stała $k_\varepsilon > 0$ taka, że jeśli *a*, *b* i *c* są liczbami naturalnymi nieposiadającymi wspólnego dzielnika z wyjątkiem 1 oraz $a + b = c$, to $c \leq k_\varepsilon P^{1+\varepsilon}$, gdzie *P* jest iloczynem wszystkich różnych dzielników pierwszych iloczynu *abc*.

98 Andrew Granville, Thomas J. Tucker, *It's as easy as* abc (*To proste jak abecadło*), „Notices of the American Mathematical Society" 2002, tom 49, s. 1224–1231.

We wrześniu 2012 roku Shinichi Mochizuki ogłosił, że udało mu się udowodnić hipotezę ABC dzięki zastosowaniu zupełnie nowego podejścia do podstaw geometrii algebraicznej. Specjaliści sprawdzają jego 500-stronicowy dowód, co zapewne jeszcze trochę potrwa.

Słownik

algorytm: Określona procedura rozwiązywania problemu, gwarantująca uzyskanie odpowiedzi.

arytmetyka modularna: System arytmetyczny, w którym wielokrotności określonej liczby, nazywanej *modułem*, traktuje się tak, jak gdyby wynosiły 0.

asymptotyczność: Mówimy, że dwie wielkości zdefiniowane za pomocą tej samej zmiennej są asymptotyczne, jeśli ich iloraz zbliża się do 1, gdy wartość zmiennej rośnie.

bozon Higgsa: Cząstka elementarna wyjaśniająca pochodzenie mas wszystkich cząstek. W lipcu 2012 roku uczeni pracujący w Wielkim Zderzaczu Hadronów ogłosili, że udało im się odkryć taką cząstkę.

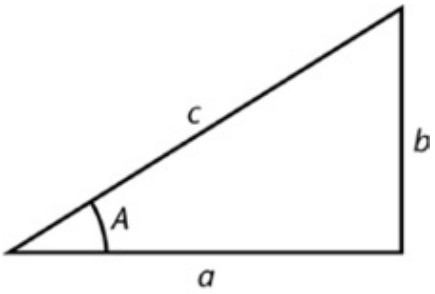
chaos: Pozornie przypadkowe zachowanie układu deterministycznego.

charakterystyka Eulera: Wzór $S - K + W$, gdzie S jest liczbą ścian triangulacji jakiejś przestrzeni, K to liczba krawędzi, W jest zaś liczbą wierzchołków. W przypadku torusa mającego g otworów charakterystyka Eulera jest równa $2 - 2g$, bez względu na triangulację.

ciąg: Lista liczb ułożonych w kolejności. Przykładem może być ciąg potęg liczby dwa: 1, 2, 4, 8, 16…

ciąg arytmetyczny: Ciąg liczb, w którym każdy kolejny wyraz jest równy poprzedniemu zwiększonemu o pewną stałą wartość, zwaną różnicą ciągu. Na przykład ciąg 2, 5, 8, 11, 14… jest ciągiem arytmetycznym o różnicy wynoszącej 3. Kiedyś ciąg taki nazywano „postępem arytmetycznym”.

cosinus: Funkcja trygonometryczna kąta, którą dla trójkąta pokazanego na rycinie 51 można zdefiniować wzorem: $\cos A = a/c$.



Ryc. 51. Cosinus (a/c), sinus (b/c) i tangens (b/a) kąta A

cykl: W topologii: formalna kombinacja pętli uzyskanych z triangulacji wraz z przypisanymi im liczbami. W geometrii algebraicznej: formalna kombinacja podrozmaitości z przypisanymi im liczbami.

czas wybuchu: Czas, po którym rozwiązanie równania różniczkowego przestaje istnieć.

cząstka punktowa: Masa skupiona w jednym punkcie.

czynnik fazowy: Liczba zespolona o module wynoszącym 1 (czyli leżąca na okręgu jednostkowym), przez którą mnoży się kwantowe funkcje falowe.

dwunastościan: Bryła o dwunastu ścianach (zob. ryc. 38).

dwunastościan rombowy: Bryła, której ściany tworzy 12 jednakowych rombów – równoległoboków o równych bokach (zob. ryc. 15).

dysk (topologiczny): Obszar na powierzchni dający się przekształcić w sposób ciągły w koło.

faza: Zob. czynnik fazowy.

fala: Zaburzenie rozchodzące się w ośrodku, którym może być ciało stałe, ciecz lub gaz, bez wywoływania w nim trwałych zmian.

funkcja: Reguła *f*, która po zastosowaniu do liczby *x* daje inną liczbę *f*(*x*). Jeśli na przykład $f(x) = \ln x$, to *f* jest funkcją logarytmiczną. Zmienna *x* może być rzeczywista lub zespolona

(wówczas oznacza się ją zwykle symbolem *z*). W ogólnym wypadku *x* oraz *f*(*x*) mogą należeć do określonych zbiorów; w szczególności może to być płaszczyzna lub przestrzeń.

funkcja dzeta: Funkcja zespolona wprowadzona przez Riemanna, opisująca liczby pierwsze w sposób analityczny. Definiuje się ją jako szereg $\zeta(s)=\frac{1}{1^s}+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\frac{1}{5^s}+\frac{1}{6^s}+\frac{1}{7^s}+\cdots$, który jest zbieżny, gdy część rzeczywista *s* jest większa od 1. Dokonując tak zwanego przedłużenia analitycznego, definicję tę można rozszerzyć na wszystkie liczby zespolone *s*, z wyjątkiem 1.

funkcja eliptyczna: Funkcja zespolona, która nie ulega zmianie, gdy do jej zmiennej dodaje się dwie niezależne liczby zespolone. Innymi słowy, funkcja o takiej własności, że $f(z)=f(z+u)=f(z+v)$, gdzie *v* nie jest rzeczywistą wielokrotnością *u* (zob. ryc. 30).

funkcja *L* Dirichleta: Uogólnienie funkcji dzeta Riemanna.

genus: Liczba otworów w powierzchni.

geometria nieeuklidesowa: Geometria alternatywna do euklidesowej, zachowująca wszystkie tradycyjne własności punktów i prostych, z wyjątkiem postulatu istnienia jednej tylko prostej przechodzącej przez dany punkt, która jest równoległa do innej prostej. Istnieją dwa rodzaje geometrii nieeuklidesowej: eliptyczna i hiperboliczna.

geometria rzutowa: Rodzaj geometrii, w której proste równoległe nie istnieją, ponieważ każde dwie proste zawsze się przecinają w jednym punkcie. Geometrię taką uzyskuje się z geometrii euklidesowej przez dodanie prostej niewłaściwej (czyli „prostej w nieskończoności").

grupa: Abstrakcyjna struktura algebraiczna, którą tworzą zbiór oraz zdefiniowane na nim działanie operujące na parach jego elementów, z tym że muszą być spełnione trzy warunki: działanie musi być łączne, musi istnieć element neutralny oraz każdy element zbioru musi mieć element odwrotny.

grupa homologii: Topologiczny niezmiennik przestrzeni, który definiuje się za pomocą pętli. Mówimy, że dwie pętle są homologiczne, jeśli ich różnicą jest brzeg topologicznego dysku.

grupa homotopii: Topologiczny niezmiennik przestrzeni, który definiuje się za pomocą pętli. Mówimy, że dwie pętle są homotopiczne, jeśli jedną z nich można przekształcić w drugą w sposób ciągły.

grupa kohomologii: Abstrakcyjna struktura algebraiczna związana z przestrzenią topologiczną, „dualna” do grupy homologii i do niej podobna.

grupa podstawowa: Grupa klas homotopii pętli w przestrzeni topologicznej z działaniem „podążaj najpierw wzdłuż pierwszej pętli, a potem wzdłuż drugiej”.

grupa trywialna: Grupa składająca się tylko z jednego elementu: elementu neutralnego.

hipersfera: Zob. trójwymiarowa hipersfera.

homologia: Zob. grupa homologii.

homotopia: Zob. grupa homotopii.

ideał: Zob. liczba idealna.

ideał pierwszy: Odpowiednik liczby pierwszej w systemach liczb algebraicznych.

iloraz: Ilorazem dwóch liczb *a* i *b* jest wynik z ich dzielenia a/b.

indeks punktu względem krzywej: Liczba mówiąca o tym, ile razy krzywa okrąża wybrany punkt w kierunku przeciwnym do ruchu wskazówek zegara.

indukcja matematyczna: Ogólna metoda dowodzenia twierdzeń dotyczących liczb naturalnych. Jeśli jakaś własność jest prawdziwa dla 0, a z jej prawdziwości dla dowolnej liczby naturalnej *n* wynika prawdziwość dla *n* + 1, to własność taka jest prawdziwa dla wszystkich liczb naturalnych.

jednoznaczność rozkładu na czynniki pierwsze: Własność rozkładu na czynniki pierwsze polegająca na tym, że każdą liczbę można zapisać jako iloczyn liczb pierwszych tylko w jeden sposób, nie biorąc pod uwagę kolejności wystąpienia czynników w iloczynie. Własność taka obowiązuje dla liczb naturalnych, ale w bardziej ogólnych systemach algebraicznych nie zawsze jest spełniona.

klasa E: Klasa algorytmów, których czas działania dla danych wejściowych o rozmiarze *n* jest proporcjonalny do *n*-tej potęgi jakiejś stałej.

klasa Hodge'a: Klasa kohomologii cykli rozmaitości algebraicznej o specjalnych własnościach analitycznych.

klasa P: Klasa algorytmów, których czas działania jest proporcjonalny do stałej potęgi rozmiaru danych wejściowych.

klasa nie-P: Klasa algorytmów nienależących do klasy P.

klasa NP: Klasa problemów, których rozwiązanie można sprawdzić (ale niekoniecznie znaleźć) za pomocą algorytmu klasy P.

kohomologia: Zob. grupa kohomologii.

konfiguracja nieunikniona: Jeden z elementów listy zawierającej konfiguracje sieci, która charakteryzuje się tym, że w dowolnej sieci na płaszczyźnie musi się znaleźć przynajmniej jeden z jej elementów.

konfiguracja redukowalna: Część sieci charakteryzująca się tym, że jeśli po jej usunięciu resztę sieci można pokolorować czterema barwami, to taką samą operację da się przeprowadzić również na sieci wyjściowej.

konstrukcja przy użyciu cyrkla i linijki (konstrukcja klasyczna): Każda konstrukcja geometryczna, którą można przeprowadzić z wykorzystaniem linijki bez podziałki oraz cyrkla.

kontrprzykład: Przykład dowodzący niepoprawności jakiegoś stwierdzenia. Na przykład liczba 9 jest kontrprzykładem stwierdzenia: „wszystkie liczby nieparzyste są liczbami pierwszymi".

krzywa eliptyczna: Krzywa na płaszczyźnie, której równanie ma postać $y^2 = ax^3 + bx^2 + cx + d$, przy czym przyjmuje się zwykle, że stałe *a*, *b*, *c* i *d* są wymierne (zob. ryc. 27).

krzywizna: Miara wygięcia przestrzeni w pobliżu danego punktu. Sfera ma dodatnią krzywiznę, płaszczyzna – zerową, natomiast przestrzeń o kształcie siodła ma krzywiznę ujemną.

kula: Sfera wypełniona w środku, czyli sfera wraz ze swoim wnętrzem.

kwadrat: Liczba pomnożona przez siebie. Na przykład kwadrat liczby 7, zapisywany jako 7^2, wynosi 7 × 7 = 49.

kwantowa funkcja falowa: Funkcja matematyczna opisująca własności układu kwantowego.

kwantowa teoria pola: Teoria kwantowomechaniczna wielkości, która wypełnia całą przestrzeń i może przyjmować różne wartości w różnych miejscach (i zwykle tak się dzieje).

liczba algebraiczna: Liczba zespolona będąca rozwiązaniem równania wielomianowego o współczynnikach całkowitych lub wymiernych (przypadki te są sobie równoważne). Przykładem może być liczba $i\sqrt{2}/3$, która jest rozwiązaniem równania $x^2+\frac{2}{9}=0$ lub, w zapisie równoważnym, $9x^2+2=0$.

liczba algebraiczna całkowita: Liczba zespolona będąca rozwiązaniem równania wielomianowego o współczynnikach całkowitych i współczynniku przy najwyższej potędze równym 1. Przykładem może być liczba $i\sqrt{2}$, która jest rozwiązaniem równania $x^2+2=0$.

liczba całkowita: Dowolna liczba z ciągu: …, –3, –2, –1, 0, 1, 2, 3, …

liczba cyklotomiczna (całkowita): Kombinacja liniowa potęg pierwiastka zespolonego z jedności o wymiernych (całkowitych) współczynnikach.

liczba Fermata: Liczba postaci $2^{2^k}+1$, gdzie *k* jest nieujemną liczbą całkowitą. Jeśli liczba ta jest liczbą pierwszą, to nazywamy ją *liczbą pierwszą Fermata*.

liczba idealna: Liczba, która nie należy do danego zbioru liczb algebraicznych, ale jest z nim związana tym, że przywraca jednoznaczność rozkładu na czynniki pierwsze w przypadkach, gdy własność ta nie jest zachowana w oryginalnym zbiorze. We współczesnej algebrze jej rolę odgrywa ideał, który jest specjalnym rodzajem podzbioru w analizowanym zbiorze liczbowym.

liczba naturalna: Dowolna liczba z ciągu: 0, 1, 2, 3, …

liczba niewymierna: Liczba rzeczywista, która nie jest liczbą wymierną, czyli innymi słowy taka, której nie da się zapisać w postaci p/q, gdzie *p* i *q* są liczbami całkowitymi, a $q \uparrow 0$. Przykładami takich liczb są $\sqrt{2}$ i π.

liczba pierwsza: Liczba naturalna większa od 1, której nie można otrzymać w wyniku pomnożenia przez siebie dwóch mniejszych liczb naturalnych. Kilka najmniejszych liczb pierwszych to: 2, 3, 5, 7, 11, 13.

liczba przestępna: Liczba, która nie jest rozwiązaniem żadnego równania algebraicznego o wymiernych współczynnikach. Przykładami mogą być liczby π oraz *e*.

liczba przystająca: Liczba, która jest wspólną różnicą ciągu trzech kwadratów liczb wymiernych.

liczba rzeczywista: Każda liczba dająca się zapisać za pomocą cyfr systemu dziesiętnego, które mogą ciągnąć się nawet w nieskończoność – na przykład: π = 3,14159265358979323 85…

liczba wymierna: Liczba rzeczywista postaci p/q, gdzie *p* i *q* są liczbami całkowitymi, a $q \uparrow 0$. Przykładem może być liczba $22/7$.

liczba zespolona: Liczba postaci $a+bi$, gdzie *i* jest pierwiastkiem kwadratowym z minus jeden, natomiast *a* oraz *b* są liczbami rzeczywistymi.

liczba złożona: Liczba naturalna, którą można otrzymać przez wymnożenie dwóch mniejszych liczb naturalnych.

logarytm całkowy: Funkcja $\mathrm{Li}(x) = \int_0^x \frac{dt}{\ln t}$.

logarytm naturalny: Logarytm naturalny z liczby *x*, symbolicznie zapisywany jako $\ln x$, jest równy potędze, do której należy podnieść *e* (= 2,71828…), aby otrzymać *x*. Innymi słowy: $e^{\ln x} = x$.

maksimum: Największa wartość jakiejś wielkości.

miejsce zerowe (funkcji): Jeśli *f* jest funkcją, to *x* jest jej miejscem zerowym wtedy, gdy $f(x) = 0$.

minimalny kontrprzykład: Obiekt matematyczny nieposiadający określonej cechy, który jest w pewnym sensie najmniejszy

z wszystkich tego typu obiektów, jakie są możliwe. Przykładem może być mapa niedająca się pokolorować czterema barwami o najmniejszej liczbie obszarów. Minimalne kontrprzykłady są często obiektami hipotetycznymi i celem argumentacji jest udowodnienie, że nie mogą istnieć.

minimum: Najmniejsza wartość jakiejś wielkości.

Model Standardowy: Model mechaniki kwantowej obejmujący wszystkie znane cząstki elementarne.

moment pędu: Miara ilości ruchu obrotowego danego ciała.

obrót: Na płaszczyźnie: przekształcenie obracające wszystkie punkty o taki sam kąt wokół ustalonego punktu nazywanego środkiem obrotu. W przestrzeni: przekształcenie obracające wszystkie punkty o taki sam kąt wokół ustalonej prostej, nazywanej osią obrotu.

ogólna teoria względności: Opracowana przez Einsteina teoria grawitacji, w której siłę grawitacyjną interpretuje się jako zakrzywienie czasoprzestrzeni.

ograniczenie górne: Określona wartość, o której wiadomo, że musi być większa od poszukiwanej wielkości.

operator: Specjalny rodzaj funkcji A, która po zastosowaniu do wektora $\boldsymbol{v}$ daje inny wektor $A\boldsymbol{v}$. Operator musi spełniać warunki przekształcenia liniowego, to znaczy muszą zachodzić równości: $A(\boldsymbol{v} + \boldsymbol{w}) = A\boldsymbol{v} + A\boldsymbol{w}$ oraz $A(a\boldsymbol{v}) = aA(\boldsymbol{v})$ dla dowolnej stałej a.

optymalizacja: Szukanie maksimum lub minimum jakiejś funkcji.

osobliwość: Punkt, w którym dzieje się coś niedobrego – na przykład funkcja przyjmuje nieskończoną wartość lub rozwiązanie jakiegoś równania przestaje istnieć.

oś obrotu: Prosta, wokół której obraca się ciało.

pęd: Iloczyn masy i prędkości.

pętla: Krzywa zamknięta w przestrzeni topologicznej.

pierwiastek z jedynki: Liczba zespolona ζ, której pewna potęga ζ^k wynosi 1 (zob. ryc. 7).

pięciokąt: Wielokąt o pięciu bokach.

płaski torus: Torus powstały w wyniku utożsamienia ze sobą przeciwległych boków kwadratu (zob. ryc. 12). Naturalna

geometria takiego torusa ma zerową krzywiznę.

pole elektromagnetyczne: Funkcja określająca wielkość i kierunek pola elektrycznego i magnetycznego w każdym punkcie przestrzeni.

pole prędkości: Funkcja określająca prędkość w każdym punkcie przestrzeni. Przykładem może być przepływ cieczy, której prędkość można określić w każdym punkcie i zazwyczaj jest ona różna w różnych miejscach.

potęga: Liczba pomnożona przez siebie określoną ilość razy. Na przykład czwarta potęga liczby 3, którą zapisujemy symbolicznie jako 3^4, wynosi $3 \times 3 \times 3 \times 3 = 81$.

powierzchnia: Kształt przestrzenny powstały w wyniku połączenia ze sobą obszarów, które z topologicznego punktu widzenia są równoważne wnętrzu koła. Przykładami mogą być sfera i torus.

prędkość: Tempo zmiany położenia z upływem czasu. Prędkość ma zarówno wartość (nazywaną szybkością), jak i kierunek.

problem NP-zupełny: Określony problem klasy NP charakteryzujący się tym, że jeśli uda się go rozwiązać za pomocą jakiegoś algorytmu klasy P, to będzie to oznaczało, że *wszystkie* problemy klasy NP można rozwiązać za pomocą algorytmów klasy P.

przekształcenie: Synonim terminu „funkcja”, którym posługujemy się zwykle wtedy, gdy zmiennymi są punkty w jakiejś przestrzeni. Przykładem przekształcenia kwadratu może być „obrót o kąt prosty wokół środka”.

przekształcenie ciągłe: Przekształcenie przestrzeni, charakteryzujące się tym, że po jego zastosowaniu punkty leżące bardzo blisko siebie nie odsuną się nagle na dużą odległość.

przepływ Ricciego: Równanie opisujące, jak krzywizna przestrzeni zmienia się z upływem czasu.

przestrzeń topologiczna: Kształt, który uważa się za „taki sam”, nawet gdy zostanie poddany dowolnemu przekształceniu ciągłemu.

przesunięcie (translacja): Przekształcenie przestrzeni, w wyniku którego wszystkie punkty przesuwają się w jednakowym kierunku

o taką samą odległość.

ranga: Największa liczba niezależnych rozwiązań wymiernych równania definiującego krzywą eliptyczną. Przez rozwiązania „niezależne” rozumiemy takie, których nie można wyprowadzić z innych rozwiązań za pomocą standardowej konstrukcji geometrycznej pozwalającej uzyskać nowe rozwiązanie z dwóch innych znanych wcześniej (zob. ryc. 25).

rozkład na czynniki pierwsze: Zapisanie liczby za pomocą iloczynu jej czynników pierwszych. Na przykład rozkładem na czynniki pierwsze liczby 60 jest iloczyn $2^2 \times 3 \times 5$.

rozmaitość topologiczna: Wielowymiarowy odpowiednik gładkiej powierzchni.

rozmaitość algebraiczna: Przestrzeń wielowymiarowa określona przez układ równań algebraicznych.

równanie diofantyczne: Równanie, którego rozwiązań szuka się w zbiorze liczb wymiernych.

równanie kwadratowe: Dowolne równanie postaci $ax^2 + bx + c = 0$, gdzie *x* jest niewiadomą, natomiast *a*, *b* i *c* to stałe.

równanie różniczkowe: Równanie wiążące funkcję z szybkością zmiany jej wartości.

równanie różniczkowe cząstkowe: Równanie różniczkowe opisujące szybkość zmiany wartości danej funkcji względem dwóch lub więcej różnych zmiennych (często są to przestrzeń i czas).

równanie sześcienne: Każde równanie postaci $ax^3 + bx^2 + cx + d = 0$, gdzie *x* jest niewiadomą, natomiast *a*, *b*, *c* oraz *d* – stałymi.

sfera: Zbiór wszystkich punktów przestrzeni znajdujących się w określonej odległości od ustalonego punktu, nazywanego środkiem sfery. Sfera jest okrągła, tak jak kula, ale tworzą ją tylko punkty znajdujące się na powierzchni kuli, z wyłączeniem jej wnętrza.

sieć: Na płaszczyźnie: układ punktów powtarzający się w dwóch niezależnych kierunkach, niczym wzór na tapecie (zob. ryc. 26).

W przestrzeni: układ punktów powtarzający się w trzech niezależnych kierunkach przypominający układ atomów w krysztale. W teorii grafów: zbiór punktów (wierzchołków, węzłów) połączony liniami (krawędziami).

sieć dualna: Sieć utworzona na podstawie innej sieci przez skojarzenie węzła z każdym jej obszarem i połączenie tak otrzymanych węzłów krawędziami, gdy odpowiadające im obszary sąsiadują ze sobą (zob. ryc. 10).

sinus: Funkcja trygonometryczna kąta, na rycinie 51 na zdefiniowana jako $\sin A = b/c$.

stała Eulera: Specjalna liczba, oznaczana symbolem γ, równa w przybliżeniu 0,57721.

stan niestabilny: Stan układu dynamicznego, do którego układ może już nie powrócić, gdy podda się go niewielkim nawet zaburzeniom.

stan stabilny: Stan układu dynamicznego, do którego powraca on, jeśli poddamy go niewielkim zaburzeniom.

stopień wielomianu: Najwyższa potęga zmiennej w wielomianie. Na przykład wielomian $6x^3 - 5x^2 + 4x - 7$ jest wielomianem stopnia 3.

symetria: Przekształcenie dowolnego obiektu, które nie zmienia jego ogólnej postaci. Przykładem może być obrót kwadratu o kąt prosty.

symetria cechowania: Grupa symetrii lokalnych układu równań. Innymi słowy, są to przekształcenia zmiennych, które mogą być różne w różnych punktach przestrzeni, charakteryzują się jednak tym, że każde rozwiązanie równań nadal pozostaje rozwiązaniem, pod warunkiem uwzględnienia odpowiednich poprawek o właściwym znaczeniu fizycznym.

szereg: Wyrażenie, w którym dodaje się do siebie wiele wielkości – często nieskończenie wiele.

szereg potęgowy: Wyrażenie podobne do wielomianu, różniące się jednak od niego tym, że w szeregu może wystąpić nieskończenie wiele potęg danej zmiennej – na przykład: $1 + 2x + 3x^2 + 4x^3 + \cdots$. W odpowiednich okolicznościach taka nieskończona suma może

mieć dobrze określoną wartość i wtedy mówimy, że szereg taki jest zbieżny.

sześcian: Liczba pomnożona trzykrotnie przez siebie. Na przykład sześcianem liczby 7 jest 7 × 7 × 7 = 343. Zazwyczaj zapisuje się go jako 7^3.

tangens: Funkcja trygonometryczna kąta, którą na rycinie 51 określono jako $\operatorname{tg} A = b/a$.

teoria cechowania: Kwantowa teoria pola z grupą symetrii cechowania.

topologia: Nauka zajmująca się badaniem przestrzeni topologicznych.

torus: Powierzchnia w kształcie obwarzanka z jednym otworem (zob. ryc. 12).

triangulacja: Pokrycie powierzchni siecią trójkątów lub ich wielowymiarowych odpowiedników.

trójka pitagorejska: Trzy liczby naturalne *a*, *b*, *c* takie, że $a^2 + b^2 = c^2$. Na przykład: *a* = 3, *b* = 4, *c* = 5. Z twierdzenia Pitagorasa wynika, że tego typu liczby są długościami boków trójkąta prostokątnego.

trójwymiarowa hipersfera: Trójwymiarowy odpowiednik sfery, czyli zbiór wszystkich punktów czterowymiarowej przestrzeni znajdujących się określonej odległości od ustalonego punktu, będącego jej środkiem.

trysekcja: Podział na trzy równe części – określenia tego używa się najczęściej w odniesieniu do kąta.

układ dynamiczny: Dowolny układ, którego stan zmienia się z upływem czasu zgodnie z określonymi regułami. Przykładem może być ruch planet w Układzie Słonecznym.

układ okresowy: Układ powtarzający bez końca to samo zachowanie.

układ regularny centrowany ściennie: Powtarzający się układ punktów w przestrzeni powstały przez ułożenie sześcianów w strukturę trójwymiarowej szachownicy, a następnie uwzględnienie wszystkich ich wierzchołków i środków każdej ze ścian (zob. ryc. 17 oraz 19).

upakowanie: Zbiór obiektów przestrzeni ułożonych tak, by na siebie nie nachodziły.

upakowanie sieciowe: Zbiór jednakowych kół lub kul, których środki tworzą sieć.

wartość własna: Jedna z liczb ze zbioru specjalnych wartości związanych z operatorem. Jeśli po zastosowaniu operatora do jakiegoś wektora otrzymujemy ten sam wektor pomnożony przez stałą, to stała ta jest wartością własną.

wektor: W mechanice: wielkość mająca wartość i kierunek. W algebrze i analizie matematycznej: uogólnienie tego pojęcia.

wielokąt: Figura płaska, której brzegiem jest skończona liczba odcinków.

wielokąt foremny: Wielokąt, którego wszystkie boki i kąty są takie same (zob. ryc. 4).

wielomian: Wyrażenie algebraiczne, takie jak $6x^3 - 5x^2 + 4x - 7$, w którym różne potęgi zmiennej *x* mnoży się przez stałe, a następnie do siebie dodaje.

wielościan: Bryła ograniczona skończoną liczbą wielokątów.

wielościan foremny: Bryła, której ścianami są jednakowe wielokąty foremne ułożone w taki sam sposób we wszystkich rogach. Euklides udowodnił, że istnieje dokładnie pięć wielościanów foremnych.

wir: Ciecz płynąca wokół określonego punktu. Wiry mogą mieć różny rozmiar – w szczególności mogą być bardzo małe.

współczynnik: W wielomianach takich jak $6x^3 - 5x^2 + 4x - 7$ współczynnikami są liczby 6, –5, 4, –7, przez które mnoży się poszczególne potęgi *x*.

współrzędna: Jedna z liczb na liście określającej położenie punktu na płaszczyźnie lub w przestrzeni.

wykładnik: Jeśli zmienna *x* jest podniesiona do jakiejś potęgi, to wykładnikiem jest właśnie owa potęga. Na przykład w wyrażeniu x^7 wykładnikiem jest liczba 7.

wymiar: Liczba współrzędnych potrzebnych do określenia położenia punktu w danej przestrzeni. Na przykład płaszczyzna ma dwa

wymiary, przestrzeń zaś, w której żyjemy (opisywana przez geometrię euklidesową), ma trzy wymiary.

zbiór: Zestaw obiektów (matematycznych). Przykładem może być zbiór wszystkich liczb całkowitych.

zmienna: Wielkość, która może przyjmować dowolną wartość z pewnego zakresu.

Literatura uzupełniająca

Książki oznaczone gwiazdką (*) wymagają wiedzy specjalistycznej.

W języku polskim

Krzysztof Ciesielski, Zdzisław Pogoda, *Bezmiar matematycznej wyobraźni*, Prószyński i S-ka, Warszawa 2005.

Jacques Hadamard, *Psychologia odkryć matematycznych*, przeł. Rafał Molski, Państwowe Wydawnictwo Naukowe, Warszawa 1964.

Leopold Infeld, *Wybrańcy bogów: powieść o życiu Ewarysta Galois*, przeł. Stanisław Sielski, Prószyński i S-ka, Warszawa 1998.

Michio Kaku, *Hiperprzestrzeń*, przeł. Ewa L. Łokas i Bogumił Bieniok, Prószyński i S-ka, Warszawa 2010.

Zdzisław Pogoda, Krzysztof Ciesielski, *Diamenty matematyki*, Prószyński i S-ka, Warszawa 1997.

Lisa Randall, *Ukryte wymiary Wszechświata*, przeł. Bogumił Bieniok i Ewa L. Łokas, Prószyński i S-ka, Warszawa 2011.

Ian Sample, *Peter Higgs*, przeł. Bogumił Bieniok i Ewa L. Łokas, Prószyński i S-ka, Warszawa 2012.

Simon Singh, *Tajemnica Fermata*, przeł. Paweł Strzelecki, Prószyński i S-ka, Warszawa 1999.

Ian Stewart, *17 równań, które zmieniły świat*, przeł. Julia Szajkowska, Prószyński i S-ka, Warszawa 2013.

Ian Stewart, *Czy Bóg gra w kości: nowa matematyka chaosu*, przeł. Włodzimierz Komar, Michał Tempczyk, Państwowe Wydawnictwo Naukowe, Warszawa 2001.

Ian Stewart, *Dlaczego prawda jest piękna*, przeł. Tomasz Krzysztoń, Prószyński i S-ka, Warszawa 2012.

Ian Stewart, *Gabinet zagadek matematycznych*, przeł. Agnieszka Sobolewska, Wydawnictwo Literackie, Kraków 2011.

Ian Stewart, *Gabinet matematycznych zagadek cz. 2*, przeł. Agnieszka Sobolewska, Wydawnictwo Literackie, Kraków 2012.

Ian Stewart, *Histerie matematyczne: gry i zabawy z matematyką*, przeł. Paweł Strzelecki, Prószyński i S-ka, Warszawa 2007.
Ian Stewart, *Jak pokroić tort i inne zagadki matematyczne*, przeł. Agnieszka Sobolewska, Prószyński i S-ka, Warszawa 2012.
Ian Stewart, *Krowy w labiryncie i inne eksploracje matematyczne*, przeł. Agnieszka Sobolewska, Prószyński i S-ka, Warszawa 2011.
Ian Stewart, *Liczby natury*, przeł. Michał Tempczyk, CIS, „Presspublica", Warszawa 1996.
Ian Stewart, *Listy do młodego matematyka*, przeł. Paweł Strzelecki, Prószyński i S-ka, Warszawa 2008.
Ian Stewart, *Matematyka życia*, przeł. Bogumił Bieniok i Ewa L. Łokas, Prószyński i S-ka, Warszawa 2014.
Ian Stewart, *Oswajanie nieskończoności*, przeł. Bogumił Bieniok i Ewa L. Łokas, Prószyński i S-ka, Warszawa 2009.
Ian Stewart, *Stąd do nieskończoności*, przeł. Jacek Bańkowski, Prószyński i S-ka, Warszawa 2012.

W języku angielskim

* Colin C. Adams, *The Knot Book* (*Księga węzłów*), W.H. Freeman, 1994.
* Felix Browder [red.], *Mathematical Developments Arising from Hilbert Problems* (*Rozwój matematyki zapoczątkowany przez problemy Hilberta*), „Proceedings of Symposia in Pure Mathematics" tom 28, American Mathematical Society, 1976.
* Tian Yu Cao, *Conceptual Developments of 20th Century Field Theories* (*Rozwój pojęciowy dwudziestowiecznych teorii pola*), Cambridge University Press, 1997.
William J. Cook, *In Pursuit of the Travelling Salesman* (*W pogoni za komiwojażerem*), Princeton University Press, 2012.
Keith Devlin, *The Millennium Problems* (*Problemy milenijne*), Granta, 2004.
Frolin Diacu, Philip Holmes, *Celestial Encounters* (*Spotkania na niebie*), Princeton University Press, 1999.
Underwood Dudley, *A Budget of Trisections* (*Zapas trysekcji*), Springer, 1987.
Underwood Dudley, *Mathematical Cranks* (*Matematyczni ekscentrycy*), Mathematical Association of America, 1992.
Marcus Du Sautoy, *The Music of Primes* (*Muzyka liczb pierwszych*), Harper Perennial, 2004.
Masha Gessen, *Perfect Rigour* (*Ścisłość naukowa*), Houghton Mifflin, 2009.

* Jay R. Goldman, *The Queen of Mathematics* (*Królowa matematyki*), A.K. Peters, 1998.
* Harris Hancock, *Lectures on the Theory of Elliptic Functions* (*Wykłady z teorii funkcji eliptycznych*), Dover, 1958.
* Jeffrey C. Lagarias, *The Ultimate Challenge: The* 3*x*+1 *Problem* (*Ostateczne wyzwanie: problem* 3*x*+1), American Mathematical Society, 2011.
* Charles Livingston, *Knot Theory* (*Teoria węzłów*), „Carus Mathematical Monographs" tom 24, Mathematical Association of America, 1993.
M. Livio, *The Equation That Couldn't Be Solved* (*Równanie, którego nie udało się rozwiązać*), Simon & Schuster, 2005.
* Henry McKean, Victor Moll, *Elliptic Curves* (*Krzywe eliptyczne*), Cambridge University Press, 1997.
Donald O'Shea, *The Poincaré Conjecture* (*Hipoteza Poincarégo*), Walker, 2007.
* Gerhard Ringel, *Map Color Theorem* (*Twierdzenie o kolorowaniu map*), Springer, 1974.
* C. Ambrose Rogers, *Packing and Covering* (*Upakowywanie i pokrywanie*), „Cambridge Tracts in Mathematics and Mathematical Physics" tom 54, Cambridge University Press, 1964.
Karl Sabbagh, *Dr Riemann's Zeros* (*Zera doktora Riemanna*), Atlantic Books, 2002.
* René Schoof, *Catalan's Conjecture* (*Hipoteza Catalana*), Springer, 2008.
George Szpiro, *Kepler's Conjecture* (*Hipoteza Keplera*), Wiley, 2003.
* Jean-Pierre Tignol, *Galois' Theory of Algebraic Equations* (*Teoria Galois równań algebraicznych*), Longman Scientific and Technical, 1980.
Matthew Watkins, *The Mystery of the Prime Numbers* (*Tajemnica liczb pierwszych*), Inamorata Press, 2010.
Robin Wilson, *Four Colours Suffice* (*Cztery kolory wystarczą*), Allen Lane, 2002.
Benjamin Yandell, *The Honors Class* (*Najlepsi z najlepszych*), A.K. Peters, 2002.